‘Berrypicking’ in History. A user-centered approach to bibliographic interfaces.

Afstudeerscriptie Informatiekunde

Peter Scholing

Scriptiebegeleider en eerste lezer:
Dr. L.M. Bosveld
Tweede lezer:
Dr. H. Ellermann

Informatiekunde
Rijks*universiteit* Groningen
2007

Acknowledgements

I would like dedicate this thesis to all those who stuck by me while I was working on (and toward) my thesis, especially my parents and family.

There are a few people I would like to mention specifically, and thank them for their contributions to this thesis:

- Dr. L.M. Bosveld, my thesis supervisor and first reader, for her good advice and her patience.
- Dr. H.H. Ellermann, director of the University of Groningen's Digital Library Facility Department and second reader, for providing me with all facilities at the library, which enabled me to work on the thesis both within and outside office hours.
- Drs. I. Fahmi, SWHi project leader, for all feedback, advice, and for our long brainstorm/discussion sessions.

Peter Scholing,
Groningen, August 31st, 2007.

Abstract

Most current-day on-line information finding aids (search engines, digital library catalogs) are geared toward the one-stop-search paradigm of the traditional Information Retrieval Model: 'one–query, one–use'. This paradigm may suit a large group of users in their basic searching needs, but for many, 'one size' may not fit all.

In this thesis we have attempted to formulate a model for an 'ideal' on-line finding aid for one such group of users, for which the 'straightforward' traditional IR model is not a perfect fit: historians. As frequent scholarly users of archives and libraries, they display a more organic approach to information-seeking. Their behavior resembles the 'Berrypicking' model as described by Bates (1989), and involves a wide range of techniques not covered by the traditional model, like serendipitous browsing, name-collecting and citation chaining.

In order to come to such a model we have not only analyzed *the user*, but also *the usage*—the historians' usage of primary resource material and bibliographic (meta)data, to be exact. For this we have examined the *Evans* corpus, a set of over 40,000 printed primary source documents concerning early American history, politics and culture. *Evans* is a premier example of the kind of bibliographic data historians come across and use in their daily research practice.[1]

We have translated our findings regarding user and data into a new proposed model— and accompanying algorhitm—for historical interface design. This model—dubbed the 'Semantically-enhanced Berry Basket'—combines the flexibility and usability of the Social Web (or Web 2.0) with the power and precision of the Semantic Web. The 'cognition-augmenting' capabilities of Information Visualizalizations (Card et al., 1999), when combined with powerful 'berry'-suggestion mechanisms (powered by 'Semantic' triples) form a promising basis for a user-/historian-centered on-line electronic finding aid for primary resource material.

Keywords

Human-Computer Interaction, Information-Seeking Behavior, Berrypicking, Automatic Similarity Suggestion, Information Visualization, Semantic Web, Web 2.0, History, Metadata, Digital Libraries, *Evans*.

[1] *Evans* is also the main focus of the SWHi (Semantic Web for History) project, the context in which the research for this thesis was performed.

Contents

Chapter 1

Introduction

The advent of the World Wide Web (the Web) has brought a myriad of new dimensions to scholarship and research. It has brought people together in offering new ways to communicate and collaborate, and in delivering information to people's doorsteps. A host of on-line information services has been developed for scholars of all fields, and the scramble for digitization has commenced: search engines can dig up almost any piece of (scholarly) information, regardless of its publishing medium (website, scholarly article, book); repositories have been set up to facilitate research and peer-review; libraries, museums and archives have started to open up their collections, either by digitizing them or by providing electronic finding aids to these collections.

But have all fields of research been served equally? Take historians, for instance. Historians have traditionally been among the most frequent users of libraries and archives. They are foremost researchers of information and documentation. However, at the same time they have been—to put it mildly—a bit wary of all things digital. This can partly be attributed to aspects of the historian's trade, one of which is a principal reliance on written sources. But can it also be due to on-line information services failing to reach this (important) audience? In developing an on-line information service, an electronic finding aid, specifically targeting historians as an intended audience, one certainly needs to look into these aspects.

1.1 User-Centered Design: The HCI Approach

In this thesis an attempt will be made to come to a set of requirements and recommendations for a usable and successful electronic finding aid for historical primary resource data. This is achieved by approaching the subject matter from various different angles, incorporating techniques and methods spanning and combining a variety of research fields: Human-Computer Interaction, Information

Visualization, Library Science and (Print) History.

Our objects of study will be the following: the users (historians), the information source (the Evans corpus[1]), and the interface between the user and the information source (information visualizations, interface elements, and the proposed system). By studying different aspects of what makes a successful electronic finding aid we attempt to achieve exactly that.

The aforementioned principles of user-centered analysis, information visualizations and effective interfaces are all derived from the field of Human-Computer Interaction (HCI). Contrary to a still widely held belief, HCI and Usability Engineering go far beyond practical aspects of (interface) design (like picking colors, discussing font sizes, and things like the wording and the placement of interface elements). In the contrary, it is almost unthinkable to try to design something successfully without keeping *users* and *usage* in mind. A product, be it a website, a toaster oven, or an alarm clock, is doomed and useless when developers disregard usability. Imagine what kind of products would come of this, when product design and development would only be led by technicians and only be aimed at the sheer technical possibilities, instead of by ease of use and what customers want! That should be the goal of any development process.

1.2 Research Questions

To achieve this goal we need to pose the question which is central to HCI: How do humans interact with computers? Or more specifically: How do historians interact with computer programs such as electronic finding aids, when seeking information for their research?

To answer this question we need to identify the major components that comprise this question:

1. ***H****uman*: Who exactly are our users? What are their key characteristics? What distinguishes historians from other users?

2. ***C****omputer* (or system): What constitutes an ideal electronic finding aid for scholarly use by historians? What is the nature of the information contained within? What are the key characteristics of the primary source material which this finding aid centers around, the *Evans* Source? What is the background of the source material? What kind of enhancements or approaches can be derived from the nature and the organization of this material? Which interface elements and system features can be used to

[1] *Evans*, or *Early American Imprints, Series I: Evans, 1639-1800*, is a premier collection of historical primary resource material. It contains the full text of all known existing books, pamphlets, and broadsides printed in the United States (or British American colonies prior to Independence) from 1639 through 1800. See also section 3.1: Evans, a History in Print.

build a finding aid to suit? And what role can Information Visualization play in such an interface?

3. ***I**nteraction*: How do historians go about finding scholarly information such as *Evans*? What tasks do they typically perform in achieving this? What kind of behavior do they display? What role does a computer (program) play in all this? How do historians use computers as information-seeking tools, and how do they value them? What kind of features/requirements can be derived from this?

1.3 Research strategy

This thesis seeks to answer these questions by employing a user-centered approach, rooted in theory and behavioral modeling, to come to an on-line electronic finding aid which implements effective information retrieval models and visualizations, supporting the user in finding information in and discovering new patterns and new pieces of relevant information within the information source.

The strategy we are adopting in this thesis is the following: First we will look at all aspects concerning the user. The next chapter seeks to answer the subquestions regarding the H(uman) and I(nteraction) parts, by means of a literature studies into theories and ethnographical studies of the Information-seeking Behavior (IB) of historians.

The C(omputer) part, which is the subject of the subsequent chapters, comprises everything regarding the computer, the system and the program: the source data and the architecture of the electronic finding aid. In the third chapter we will take an in-depth look at the source data for this system, the *Evans* set of historical primary source material. We will look into the history and the various inceptions of the *Evans* corpus, to gain understanding of the history of usage of the source. By looking through the eyes of others—bibliographers, archivists (makers), and historians (users)—we will be able to recognize and identify key traits of this resource. The structure in which bibliographers have organized the *Evans* corpus in the past will help us identify key elements and items of interest. These items can be used as pivotal points which we can use in transforming *Evans* from its raw (analog) form into electronic data formats.

This electronic data can serve as the building blocks for the interface elements, and information visualizations, which are the subject of the fourth chapter. This chapter will center on analyzing and exploring historical interface design and information visualizations.

The analyses from these three chapters will result in a set of requirements, which will be subsequently translated into a new (adapted) model for the IB of historians in an on-line finding aid environment. This model will be brought forward in the fourth chapter, along with—as a proof of concept—a proposed algorithm and an accompanying system. This proposed system will be in the form of a Rich Internet Application (RIA), an advanced on-line electronic finding and retrieval aid for digitized primary sources. With this electronic finding aid we aim to enhance and facilitate the historian's information-seeking behavior, context-building, and pattern discovery within the source/corpus. Or as Yakel (2005) describes it we aim to enable and aid users to 'search for, select, and identify the most appropriate records on their own'.

1.4 Special Focus and Deliverables

A special focus throughout his thesis will be given to the application of Semantic Web technologies the on-line electronic finding aid environment described above. This approach is motivated by the background of the SWHi Project[2], which has the goal of researching and experimenting with the application—and the implications—of Semantic Web technology in a Digital Library environment.[3] This thesis is a part of this ongoing project. Therefore the deliverables for this thesis will consist of a number of semi-functional prototypes (in the form of visual elements that can together form the interface for a semantic on-line information-seeking aid), as well as a set of requirements for a final version of such a system, and a set of recommendations for further research.

[2] SWHi stands for the Semantic Web for History. The SWHi project—led by Dr. Henk Ellerman and Drs. Ismail Fahmim—is a part of the University of Groningen's Digital Library department.

[3] The Semantic Web is an extension of the World Wide Web—a vision of a future Web—in which web content can not only be read and understood by human agents (as is the case now, with 'semi-structured' hypertext and natural language), but also in a more structured and formalized form that can be interpreted and shared by software agents, thus allowing for automated processing and interoperation between agents (Berners-Lee et al., 2001; W3C, 2001). It was originally conceived by Tim Berners-Lee, inventor of the World Wide Web and current W3C(`http://www.w3.org`) Director, who had a vision of the Web as a universal medium for data, information, and knowledge exchange (Berners-Lee, 2003). More in-depth information on the Semantic Web and the SWHi ontology can be found in Junte Zhang's excellent thesis on this subject Zhang (2006).

Chapter 2

Users: Historians and their Information-seeking Behavior

> Human beings are delightfully unpredictable, intuitive, inspirational and sometimes maddingly irrational. We need to remember that when we design (Faulkner, 2000).

Finding answers to the questions posed in the introductory chapter requires gaining insight into the user's behavior: how *users* act—or perform *tasks*—in a specific *environment*. These three aspects—users, tasks, and environment, form the basic foundations of Human Computer Interaction (HCI): Knowing User, Task, and Environment[1](Faulkner, 2000).

The goal of this chapter will be to try to analyze the *H*(uman) and *I*(nteraction) components of HCI: to try to understand the historian as a user of information resources.

The focus in this analysis will be on a literature study, in which we will investigate the questions posed in the introduction from a user perspective: Who is the historian, what are his key traits, and what behavior does he display when doing research, when interacting with information sources? In short: *What constitutes the historian's Information-Seeking behavior?*

Information-seeking behavior—sometimes referred to as information behavior, or IB for short—can be defined as 'how people need, seek, manage, give and use information in different contexts' (Fisher et al., 2005). If we have a clear picture of the prospected users of our system, and of their behavior in using the system, we will be able to define a set of requirements for making a better product: an on-line electronic finding aid that better suits the target audience.

[1] In order to develop a usable system, Xristine Faulkner (2000) proposes to use the basic tenet for usability engineering and Human Computer Interaction (HCI): 'Know the User! Know the Task!', to which she has added a third requirements: 'Know the Environment!'. (Faulkner, 2000)

2.1 Know the User: H for Historian

Users are the key ingredient to a successful product. A system that can do the most amazing things, but that cannot be operated by the people who were intended to use them, is near worthless: it should be *useful*, *usable* and *user-friendly* in order to be effective and successful.'Know the user' is—not coincidentally—the first of Faulkner's three HCI requirements (Faulkner, 2000).

2.1.1 User Profiling

In line with common HCI practice (amongst others Shneiderman (1997, 1998) and Hacknos and Redish (1998)) we will first formally *identify* and *classify* our primary audience. We identify our primary user base as consisting of history scholars: professional historians. These history scholars are typically employed by universities, colleges and other (research) institutions (libraries, museums, government, et cetera).[2] The nature of the resource allows for an even further narrowing of the intended audience to those professional historians interested in the specific domain of early American history, culture, and politics.

User characteristics

A number of classifications can be made in terms of user characteristics. Based on the classifications of Shneiderman (1998), Faulkner (2000) and of Hacknos and Redish (1998)[3], we will build a profile of 'the historian' based on the following variables: age, gender, level of education, level of expertise, and end-user class.[4]

As our primary audience consists of professional historians, we can quickly do away with most of the profile variables—age, sex, education, and language skills—as being a history scholar requires a minimal level of education, experience and skills. A typical professional historian has at least a Master's degree in History, but more commonly a PhD. This brings with it a certain level of domain expertise and experience on the job, but it also implies a certain age range (which is typically in the range of at least the mid twenties ranging up to seventy years of

[2] Other audiences—potential secondary users—can be identified as comprising of students of history (PhD and graduate), scholars of English literature and language, and of early American society and culture. We can also define a tertiary audience: amateur historians and other individuals interested in this particular slice of history. For this particular system we are focusing on our primary users, and assume that a system designed for this primary audience is also usable for these broader audiences with historical interests.

[3] Shneiderman breaks his user profiling down to five parts, 'profiles of their age, sex, physical abilities, education, and personality' (Shneiderman, 1998). Hacknos and Redish (1998) add experience on the job, language skills and geographical location to this list.

[4] End-user classes can best be described as 'a subset of the total population of end-users' (Faulkner, 2000). Four classes of end-users can be identified based on their relationship with the system and the nature of the tasks they perform: direct users, indirect users, remote users and support users.

age). Sex (gender) and physical abilities are not of particular concern here, and we assume them to be no different from general audiences.[5] As for language skills, the nature of the source and subject area—Early American History—also imply that scholars in this field have an ample understanding of the English language.

Location is a special case for historical resources. For a historian, ready access to the source material is paramount (Delgadillo and Lynch, 1999; Duff and Johnson, 2002; Stieg-Dalton and Charnigo, 2004; Cole, 2000a). In the past, historians doing research in a specific subject area either had to restrict their research to what could be studied locally, or be prepared to travel to get to the physical location of the source or the archive, and study it over there or make personal copies of the source (Cole, 2000a). With the advent of the Web, and on-line finding aids and digitized resources becoming readily available to anyone with a modern computer and an Internet connection, the geographical location of a scholar has become less and less an issue.

User Classification

To determine the level of interaction the user will have with the system, we need to determine into which *user class* historians fall as users of electronic finding aids: direct users, indirect users, remote users and support users (Faulkner, 2000). Direct users 'use the system themselves in order to carry out their duties' (Faulkner, 2000), and this is the case with the nearly all website visitors and users of electronic finding aids. Furthermore, historians tend to work alone to do their research—history is a 'solitary pursuit' (Orbach, 1991; Delgadillo and Lynch, 1999): they clearly do not fit within the other three, more 'detached', user classes (indirect users, remote users and support users).

Levels of expertise

An other classification, into levels of expertise, is a somewhat more difficult undertaking. This is partly because there is no 'single level of expertise', instead there is a number of dimensions to which the 'level of expertise' can be applied. For instance, in the context of scholars using electronic finding aids during their research these three *types of expertise* are relevant: domain expertise, computer skills, and experience in using (similar) electronic finding aids.

Furthermore, some researchers say that—even within every *type* of expertise—'there are no experts': everyone specializes in the specific task they usually perform within the system, and can therefore be considered an 'expert' in that specific area, and a 'novice' in another (Faulkner, 2000).[6] This implies that making

[5] Throughout this thesis, we will refer to the user as 'he' or 'him', where of course 'he/she' is meant. This is done only for matters of readability and convenience.

[6] There are commonly three levels to distinguish between: novice, intermediate (or intermit-

a clear distinction between different levels of expertise is impossible and therefore pointless. In our opinion this is not the case; we merely need to make a more careful taxonomy of the different types and different levels.

We consider our users to be experts in their field of work, with sufficient to excellent domain knowledge (Case, 1991b; Duff and Johnson, 2002; Cole, 2000a). Experienced historians—having done a fair amount of research—possess a large domain expertise and a vast amount of research experience. This experience plays an important role in their IB, as will be shown later.

Separate from domain knowledge is knowledge of and experience with computers, experience with similar/previous versions of the system (Hacknos and Redish, 1998)—other electronic finding aids and on-line resources like the one we are describing in this document. These are all takes on what Hacknos and Redish (1998) call 'experience on the job'.

With our system—a new on-line electronic finding aid (i.e. a new website)—defining the level of expertise the users have with the *current system* is easy: the users will be novices: 'all new users of a product are novices at first' Hacknos and Redish (1998). Therefore, provisions must be made to accommodate these novice users. These provisions gain even more importance with websites: usage is discretionary, non-mandatory. A website continuously has an influx of new, first-time users, with no prior knowledge of the particular workings of that one specific website—they are not mandatory users, as there are are a myriad of other potential information sources out there which to choose from (Nielsen, 2003)—so the vast majority of users will not be willing to spend as much time to familiarize themselves to non-standard or other new features that have a steep learning curve—how ingenious these new features may be—as they would with the features of a mandatory system (Nielsen, 1993; Faulkner, 2000).

If the users of a website can not instantly see the usefulness of a certain and the added value of a certain feature, they will simply stop using this resource and use another resource instead. As with other non-mandatory systems designed for a more general public, such systems 'will have to be usable at once or no effort will be expended to use them' (Faulkner, 2000).

2.1.2 Personality: Key Traits

The characteristic Shneiderman (1998) calls 'personality'—the user's key traits and behavior—is a subject that is so important to this study that it deserves to be looked into in great detail: this is where we find out what seperates the historian from other groups of (scholarly) users. This defines the set of requirements we use to build the proposed system.

tent) and expert. Every level has its own characteristics, and therefore its own set of needs and expectations of the system (Faulkner, 2000).

A review of literature shows a number of key traits of historians, some specific only to historians, other ones shared with scholars from related fields of research, like the humanities and the social sciences. Unlike scholars from many other fields—especially the 'hard sciences'—, historians rely much less on secondary sources and experiments for their research. Historical research consists of research of the past through primary resources. On these primary resources they attempt to explain past historical events, form theories.

The Importance of Primary Sources

Historical tradition and methodology dictates that primary sources[7]—original material closer to the events at hand, mostly textual—should be preferred to secondary sources, which are interpretations of these historical events, and therefore subject to discussion (Case, 1991a; Delgadillo and Lynch, 1999; Duff and Johnson, 2002; Cole, 2000a). Primary resources are where a historian can find proof and evidence. This tradition can be traced at least to the days of the great German historian Leopold von Ranke (1795-1886), who insisted that legitimate history should be composed from written primary sources stored in archives, which he called 'museums of written evidence' (Staley, 2003). This heavy reliance on primary—mostly textual—material makes historians the main most avid users of archives and libraries. As the past is the object of study, nearly all material predates the digital era. This may seem as a statement of the obvious, but this has serious implications when investigating the stance of historians toward digital technology in general, their (dis)trust of digital resources and of advanced computer applications, including advanced non-textual/graphical representations. Staley (2003) reflects on this as historians thinking through 'prose', through text—which is one-dimensional—instead of through complex data, which is usually multidimensional, and therefore more suited to visualize.

Attitude Toward Computers

The attitude of historians—as of most humanists—toward computers as aids in research has mostly been one of reluctance. This can be dated back for the most part to when computers were first found their way into scholarly research:

> The computer did not introduce a new way of thinking to the sciences. Formalizing theories in order to able to test them and applying

[7] The UCLA Institute on Primary Resources defines primary (re)sources as follows: 'Primary resources provide firsthand evidence of historical events. They are generally unpublished materials such as manuscripts, photographs, maps, artifacts, audio and video recordings, oral histories, postcards, and posters. In some instances, published materials can also be viewed as primary materials for the period in which they were written. In contrast, secondary materials, such as textbooks, synthesize and interpret primary materials.' `http://ipr.ues.gseis.ucla.edu/info/definition.html`

> statistics for measuring test-results has been scientific practice for a long time. Humanists on the other hand had little use for complicated equipment for their studies: a sharp pencil, some paper, a good archive, and a well-filled library were their only prerequisites (Staley, 2003).

Analogous to this difference in research paradigm between historians and other scholars, this quoute also denotes a major point in the treatment of data and information: the historian uses the information he finds in primary sources as the object of research, the information is itself the source that needs to be interpreted to formulate a theory.

Welling (1998) gives two types of reasons for the relative unpopularity or initial non-adoption of computers in historical research: lack of training being one, cultural reasons being the other. As for cultural reasons he mentions the lack of tradition of (experimental) testing—as is the case in the sciences—and the existence of a pecking order to computerized gathering of data: interpretation is esteemed much more than gathering. Historians who use computers extensively are considered as being mere 'stamp-collectors', as computers were initially used most for gathering (quantitative) data. The lack of training can be partly due to the cultural reasons mentioned before, but also due to the fact that history, as a craft, has been conducted for centuries *without* requiring the use of computers, and for the most part can be done without using them in this day and age.

Historians have been relatively conservative and traditional in their research practices, as can be expected from scholars who look at the past for their scholarship. And the ones who teach and instill these research methods in new generations are for the most part historians that have learned the craft of historical research before the digital era. The fact that historians can do without it accounts for most of the historical reluctance toward—and aptness in—the use of computers as a research tool. But, truth be told, historians nowadays have become avid users of computers and certainly have reaped the benefits of digitization and on-line information sources, and among them are many that are themselves involved in digitization and the building and management of on-line resources.

2.2 Information-Seeking Behavior: Historians interacting with their environment

After having learned more about who the users are, the next question that we should be asking, is 'What do our users do?'. What are their goals when using a resource like *Evans*, and how do they behave–how do they go about carrying out their tasks to achieve these goals? Faulkner (2000) identifies two main models to determine tasks, the first being a more formal traditional approach, by adhering to Donald Norman's Action Cycle Model Norman (1988)'; the second being a newer, more ‘controversial’ (Faulkner, 2000) approach in the context of Human-Computer Interaction, which utilizes methods provided by and derived from ethnography, sociology, and anthropology: the ethnographic approach.[8] In this thesis, a alternative approach, a dual one, is adopted—on one hand focusing on theories and models of Information-seeking behavior, on the other hand focusing on a the historian in its environment and context, based on previous studies and on a review of relevant literature, a more ‘ethnographical–by–proxy’– or ‘socio–cognitive’ approach (Bates, 2005b)—which will—combined together—lead to a formulation of the historian’s tasks and an understanding of the environment: the context in which the tasks are carried out by the user, the historian.

In any case—regardless of which methods one uses—it is good to have a good definition of the terminology to refer to. Faulkner (2000) makes use of Norman (1988)’s formal definitions of goals, tasks, and actions, as will we.

Defining Goals, Tasks and Actions

The *goal* is the state that the human wishes to achieve. The *task* is the activity required in order to bring about the state the human wishes to achieve (the goal). The *action* is the physical interaction with the system in order to carry out the user’s goal.

2.2.1 IR and IB: theories and models

A multitude of models (and theories)[9] exist on the subject of Information Retrieval and Information-seeking. Only few, however, come close to describing the information-seeking as it exists in the real world, how it is being done in reality by scholarly users looking for information in archives and libraries. One of these is the much-cited ‘Berrypicking’ model by Bates (1989). However, most on-line finding aids, such as search engines, as well as most archival and library finding aids are based on an entirely different model and Information Retrieval paradigm, namely what is commonly referred to as the ‘standard’ or ‘traditional’ model of Information Retrieval (Rasmussen, 1999; Fox and Sornil, 1999; Arms, 2000). In

[8] The ethnographical approach ‘can be seen as a means of developing the sort of knowledge about the users, their working practices and the environment in which they are operating that is needed to develop relevant systems’ (Faulkner, 2000).

[9] It is best to clarify the distinction between theories and models. For reasons of practicality we follow the definitions as Bates uses in her introductory chapter to *Theories of Information Behavior* (Bates, 2005b): A *theory* is ‘a system of assumptions, principles, and relationships posited to explain a specified set of phenomena. Theories often carry with them an implicit metatheory and methodology.’ A *model* is ‘a tentative ideational structure used as a testing device.’ Bates adds the following to these definition: ‘Most of ”‘Theory”’ in LIS is really still at the modeling stage.’

this section, both models are discussed, and perhaps even more important, the differences between them.

It is perhaps best, at this point, to make the distinction clear between Information-*seeking* on one hand, and Information *Retrieval* (IR) on the other: Information-*seeking* research focuses on 'the complex factors involved in human efforts to find information', whereas Information *Retrieval* concentrates on 'testing and improving computer retrieval algorithms', which often do not involve actual users and real-life search queries (Bates, 2005a).

The traditional Information Retrieval model

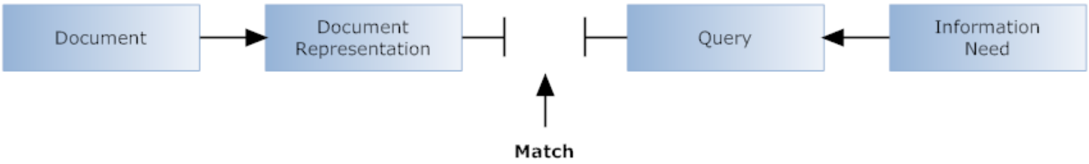


Figure 2.1: The traditional Information Retrieval model. The focus in the traditional model is the match between the document and query representations, 'one–query, one–use'. Adapted from Bates (1989).

In traditional Information Retrieval modeling and research the assumption is made that, in order to retrieve information, a user formulates a query (using keywords) and gets a set of retrieved results (Hearst, 1999). Bates (2005a) refers this as the 'one–query/one–use'–paradigm, or the 'one–stop model' (see figure 2.1). Over the years, this idea has been supplemented with the notion of 'iterative feedback', in which the retrieved set is evaluated and refined, a process is which is reiterated until a perfect result set is found.

Reformulating this with HCI terminology, the goal is a final, ultimate set of *all and only* those documents that satisfy the user's information need. This goal is achieved by a process that can be deconstructed into a cycle of eight tasks, as shown in the algorithm in table 2.1.

1) Start with an information need
2) Select a system and collections to search on
3) Formulate a query
4) Send the query to the system
5) Receive the results in the form of information items
6) Scan, evaluate, and interpret the results
7) Either stop, or
8) Reformulate the query and go to step 4

Table 2.1: The eight step cycle of tradtional Information Retrieval. Source: Hearst (1999)

The traditional focus of Information Retrieval has always been on searching—

or ad-hoc retrieval[10]—instead of on browsing (Rasmussen, 1999; Baeza-Yates and Ribeiro-Neto, 1999).

This traditional model of information retrieval is the basis for most (archival) information systems in use today (Rasmussen, 1999). This is also the primary means of locating information on the Web—by using a search engine (Baeza-Yates and Ribeiro-Neto, 1999; Bates, 2002). It is well liked for its simplicity (Bates, 1989) by system developers—because its algorithm is relatively easy to implement—and by Information Retrieval scholars—because its metrics are easy to quantify. Its 'all and only' requirement translates to the recall and precision metrics that can be used to determine and calculate the 'relevance' of the documents in the final result set to the stated query. It also serves as the basis of many theories and models: one of the most-frequently used and cited information-seeking paradigms, Ben Shneiderman's Visual Information-Seeking Mantra, 'Overview first, zoom and filter, then details-on-demand', also makes use of the traditional IR model as its underlying foundation (Shneiderman, 1996).

But, as is mentioned by many researchers, this model, deeply rooted in Information Retrieval theory, has many shortcomings. It does not seem to conform to real practice, to the way many users would like to seek information: 'real-life searches frequently do not work this way' (Bates, 1989).

Studies show that there is a clear division in the scholarly community between scholars from the 'hard' sciences, who prefer this kind of searching by query formulation, and scholars from the humanities and social sciences, who show a preference for more informal methods of information-seeking (Bass and Rosenzweig, 2001; Cole, 2000a; Bates, 2005a; Duff and Johnson, 2002; Yakel, 2005).

This model also does not take into account the fact that many users dislike being confronted with a long disorganized list of retrieval results that do not directly address their information needs (Hearst, 1999).

Bates' Berrypicking model

In the 1970s and 1980s more and more research was appearing that demonstrated how people really search for information, and especially for social scientists and humanities scholars the traditional paradigm of 'one–query/one–use' didn't quite seem to reflect actual use. As information technology progressed during the 1980s, the traditional electronic library catalogs, which were operated by trained reference librarians—who were trained to think in and make use of keywords and controlled vocabulary for actual one–stop-searching—as intermediaries, made way for on-line library catalogs and finding aids which could be used directly

[10] Keyword searching, or 'ad-hoc retrieval', as Baeza-Yates and Ribeiro-Neto (1999) call it, is defined as the 'standard retrieval task in which the user specifies his information need through a query which initiates a search (executed by the information system) for documents which are likely to be relevant to the user' (Baeza-Yates and Ribeiro-Neto, 1999)

by the end-users. However, the assumption—one–query/one–use—remained the same, as did the system design (Bates, 2005a).

In 1989, Marcia Bates came up with an alternative for the traditional IR model. She dubbed this alternative the 'Berrypicking' model. The intention of this model was to be 'much closer to the real behavior of information searchers than the traditional model of information retrieval is, and, consequently will guide our thinking better in the design of effective interfaces' (Bates, 1989). Berrypicking is named after the process of picking blueberries or huckleberries from bushes, in the forest. The berries are scattered on the bushes; they do not grow in bunches, and can therefore only be picked one at a time.

This one-at-a-time collecting of interesting pieces of information is the basis of the Berrypicking model. Another important foundation in this model is the notion of 'evolving search', which is not a part of the metaphor of picking-berries-in-the-forest, but plays an important role in Bates' model nonetheless (Bates, 1989).

The underlying idea is that end users do not have preconceived ideas about of a formulated query and clear expectations of an ultimate result set, expectations that do not change throughout the process. Instead, every bit of information a user comes across—every berry, so to speak, which can be a document, a query result, or any other piece of information in context—leads to a better conception of the domain being (re)searched—it builds the user's internal, mental conception of the information in the context of his research—which may lead to a reformulation of the goal or query, and lead the user to another direction in his quest for information within that domain. Card et al. (1999) call this sensemaking process 'knowledge crystallization': the user tries to make sense of new-found information, to see where it fits within his internal mental 'schema', and forms a new action/direction based on this. Duff and Johnson (2002) call it 'building contextual knowledge'.

Tasks in Berrypicking

Berrypicking is described as an informal way of information-collection, as opposed to the formal way, as exemplified by the 8-step task cycle of the traditional model (of table 2.1), which supposes a formal query declaration as the first step and a non-changing ultimate result set. The berrypicking process is demonstrated in figure 2.2.

End users may begin with just one feature of a broader topic, or just start off with one relevant resource, to familiarize themselves with the resource, and from there move through, employing a strategy that relies on discovery, stumbling upon new pieces of information, almost as if by accident, by serendipity.[11]

[11] Webster's on-line dictionary describes serendipity as 'the faculty of making fortunate dis-

The user can accomplish this using a variety of techniques: hierarchical browsing [12], searching, following cross-links.[13] Each new piece of information they encounter—the berry—gives them new ideas and directions to follow—adding to the 'information in their head', expanding domain knowledge and building context—and, consequently, a new conception of the query. At each stage they are not just modifying the search terms used in order to get a better match for a single query. The information-seeking goal is continually taking on new directions, is continually being 'reframed and refined'. Duff and Johnson (2002) This is what Bates calls an evolving search (Bates, 1989).

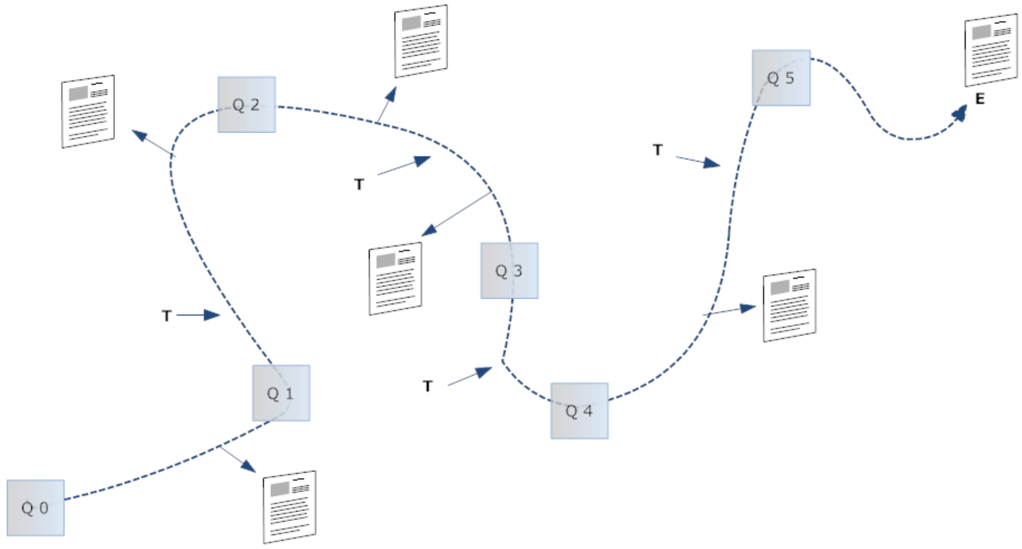


Figure 2.2: Bates' Berrypicking Information Retrieval Model. The focus is the sequence of searcher behaviours, which continually shift after gaining new information. Q = Query, T = Throught, E = Exit. Every new piece of evidence, of information—here in the form of documents—leads to reformulation of the search direction (Q) and to a new conception of the domain, of thought (T). Adapted from Bates (1989).

2.2.2 The historian's Information-seeking Behavior (HIB)

When reviewing literature on the subject of information-seeking behavior of historians, one thing sticks out: Historians also display a different information-seeking behavior than what seems to be the norm in Information Retrieval system design. They seem to closely follow the patterns laid out by Bates (1989) in her Berrypicking model, much in line with the humanists and social scientists, who also display

coveries of things you were not looking for'

[12] Browsing and its relation to searching are best described in this context by Arms' definition, who describes browsing as the 'exploration of a body of information, based on the organization of the collections or scanning lists, rather than by direct searching' (Arms, 2000).

[13] It is important to note here that the process of Berrypicking is not restricted to browsing only—as opposed to the traditional model being founded on formulation of query-by-keywords—although it may seem to be the case at first sight.

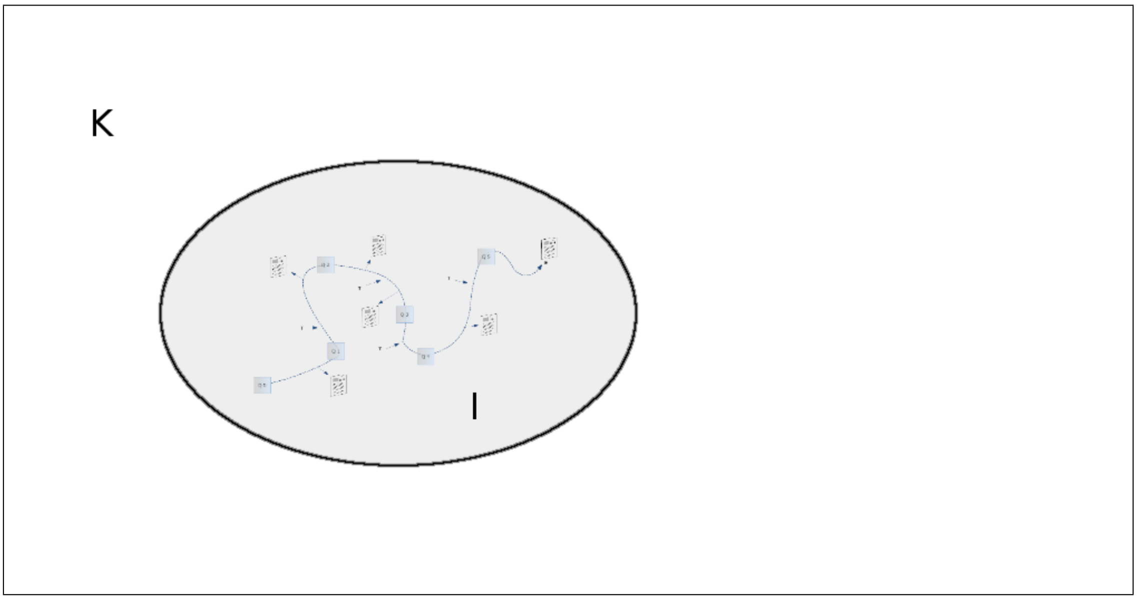


Figure 2.3: Context of Berrypicking Search. Here we see the size of the previous figure shrunk in order to show the context within which the search takes place. K = Universe of Knowledge, I = Universe of Interest. Adapted from Bates (1989).

a more informal, non-linear, type of information-seeking and -gathering behavior (Bass and Rosenzweig, 2001; Cole, 2000a; Bates, 2005a; Duff and Johnson, 2002; Yakel, 2005).

Duff and Johnson (2002) identify four stages, or four different types of information-seeking activities historians employ when doing research, which form a combination of what HCI would call goals and tasks. These four activities are:

1. Orienting oneself to archives, finding aids, sources, or a collection;
2. Seeking known material;
3. Building contextual knowledge; and
4. Identifying relevant (new) material.

Orientation and familiarization

The first stage of information-seeking is *orientation*. When orienting themselves to new archives, new finding aids and new collections, historians in general like to familiarize themselves with how the collection is organized, and try to get a clear overview by identifying common themes, important documents and names.[14] In the real (off-line) world, the initial step, when historians first visit a new archive,

[14] Rivadeneira et al. (2007) call this ‘impression formation’. Pandit and Olston (2007) refer to this as ‘orienteering’

is usually to consult with and talk to the archivist or the subject specialist in charge, to 'pick their minds' (Duff and Johnson, 2002; Delgadillo and Lynch, 1999; Tibbo, 2002). Other methods mentioned are familiarization by flipping through the index pages of the (on- or off-line) finding aid and browsing the collections' bookshelves for interesting starting points (Duff and Johnson, 2002). If this is not possible, as is sometimes the case with on-line collections, a new collection or an on-line resource can become 'overwhelming'. Researchers can experience 'a momentary sense of panic before they become oriented and are able to develop a research strategy', or a sense of 'lostness' (Duff and Johnson, 2002).

Historians seem to prefer informal methods like browsing and citation chaining to the more formal keyword searching, as their goal is to expand the 'information in their heads', the building of context (Duff and Johnson, 2002; Stieg-Dalton and Charnigo, 2004; Yakel and Torres, 2003; Bates, 2005a). In using these methods, they do not always seem to follow a singular, straightforward path, as Bass and Rosenzweig (2001) call it. Some mention the similarity of the historians' behavior to that of humanists and social scientists, and subsequently to the behavior described in Bates' Berrypicking model (Yakel, 2005; Duff and Johnson, 2002; Bates, 2005a; Bass and Rosenzweig, 2001). In the interviews conducted by Duff and Johnson (2002), historians describe this process as 'rummaging around looking for things', 'fumbling my way through this material', 'going fishing for information, hoping to catch something' and 'searching for a needle in a haystack' (Duff and Johnson, 2002). Contrary to what these descriptions may imply, the informal, serendipitous behavior displayed by historians is less haphazard and irrational than it seems: 'what appears to be accidental discovery is accidentally found on purpose' (Duff and Johnson, 2002).

Seeking Known Material: The Importance of Names

An interesting behavioral trait which distinguishes historians from other scholars is that their information-seeking behavior seems to revolve around context-building by way of name-collecting. In order to familiarize themselves to new collections or archives, historians use names as entry points.

Duff and Johnson (2002) call this the 'provenance' method, and mention common names as the most-used entry point into a collection, 'the most common element used for almost all types of queries, followed by date, place, event, and subject'. These common names are part of the strategy Shneiderman (1996) calls a 'known item search', and what Duff and Johnson (2002) refer to as 'seeking known material'.

Cole provides an example of the common practice when gathering information from archives:

> A common practice [. . .] was to collect names of people from the era and geographical era they were studying, then to place data about

> these names on [...] index cards. [...] This practice involves noting the name of an individual or a company each time that they came across it in the research material. By scan reading they are then able to slice through new material and focus on that name whenever it appears in the text. Through this methods, relationships between individual companies or persons and events become clear and patterns emerge. Historians use name collection as a cognitive tool for storing and retrieving a large quantity of individual facts. [...] Proper names serve as a mental schema for organizing and sifting through vast amounts of acquired information (Cole, 2000a).

Cole (2000a), Duff and Johnson (2002), and others mention that this preference is formed by a practice that can be traced back to the days when the only way of finding a specific document in an archive was to follow the archive's organization and setup. Archives were usually organized according to 'provenance': documents were indexed by names of their respective creators: individuals or organizations (Cole, 2000b,a; Duff and Johnson, 2002; Beattie, 1990; Orbach, 1991). Names also have another advantage over subject-based keywords, one that is especially important when dealing with historical documents: proper names have the advantage of being relatively unambiguous, whereas regular keywords describing concepts and subject terms can be ambiguous, and their meaning can change over time.

Some researchers duly note that, although this name-collecting is a commonly displayed behavioral trait, this name-collecting was imposed rather than chosen voluntarily. When given multiple entry points and a variety of information-seeking and collecting methods, the historian will adapt and incorporate these into his information-seeking practice (Beattie, 1990; Orbach, 1991; Duff and Johnson, 2002).

Perhaps the process a historian goes through when seeking information can best be illustrated along the lines of an example provided by one of the respondents in Cole's study, who, in browsing through a London-based archive in search of evidence for commercial raiding, privateering, and piracy by the English during their wars with France during Louis XIV's reign, describes his information-seeking process as a sequence of actions and realizations: (Cole, 2000a)

The researcher:

1. Compiles a list of names of ships and its captains;

2. Goes back and reformulate his research query to include the ship's owners as well;

3. Discovers an anomaly, a pattern or split in the pattern: 'discovering that there's something not right here';

4. Realizes the pattern's full nature;

5. Reformulates/refocuses research to reflect this new-found pattern, basing the new search on the new data.

Building Contextual Knowledge

Proper names not only serve as entry points into a collection. They also serve as the pivotal points on which historians build their contextual knowledge: they serve as the nodes in the network/web of knowledge, connecting the individual dots formed by the 'berries' and the proper names.

Cole's example clearly demonstrates that the historian's information-seeking centers around two notions that are important to achieve their goals: name-collecting and pattern discovery.

Along with this pattern discovery, another key concept in information-seeking is building contextual knowledge (Duff and Johnson, 2002). Historians build contextual knowledge by means of incidental discovery. They rely on informal information-seeking techniques, centered around common names, to discover new patterns. With these newly discovered patterns they augment their contextual knowledge, leading to the discovery and recognition of even more names and patterns.

Contextual knowledge is the same notion that Norman (1988) calls 'information in the head', what Card et al. (1999) and Cole (2000a) refer to as an internal or mental 'schema'. No matter what one calls it, it plays a vital role in the historian's information-seeking behavior (Cole, 2000a; Duff and Johnson, 2002). The 'schema' or 'internal context' is the internal representational framework or conceptualization of a specific problem or research question that a human forms in his mind (Card et al., 1999; Cole, 2000a). It is based on the—broader—domain knowledge and previous experience of the person, and it is augmented and fed by new pieces of information the person encounters. It can be seen as the mental picture the information-seeker, the historian in our case, forms of the specific topic he/she is researching.

An experienced historian's grasp of their knowledge domain is extensive. Because of their research experience—he can put a newly found piece of information in a context relatively easily, and a seasoned historian—making use of the 'information in his head'—can quickly assess whether that new piece of information is relevant for his research, i.e. whether it fits within the context (Duff and Johnson, 2002; Cole, 2000a). Less experienced historians, who do not (yet) possess this domain expertise and sense of context, will not be able to make these important judgments as readily (Cole, 2000a). They can, however, use the 'berries' and common names they possess knowledge of to start building their own mental framework (Cole, 2000a). In this, a system catering to this 'special' information need will be very beneficial to experienced and inexperienced historians alike.

2.3 Formalizing Goals, Tasks and Requirements

Information-gathering and context-acquiring by discovery can sometimes be hard to formalize, as ‘people do not always behave in a logical fashion’ (Norman, 1988). This sometimes irrational behavior seems also to be present with our target user group, as the literature on the information-seeking behavior of historians suggests. This behavior is rooted in the historian’s background and is—at least in part—formed by tradition and culture. Analysis on the goals and tasks historians form and carry out and the behavior they display seems to go alongside best with a more ethnographical approach.

2.3.1 The Goal: finding relevant information

A historian’s goal when using an on-line electronic finding aid, and the main reason why any user would visiting a resource like ours, is in short—as the name ‘finding aid’ suggests—to seek and find information which can help him in his research. He wants to identify all pieces of information he deems relevant to his research. This ‘relevancy’ is a very broad notion, and should be perhaps be elaborated on a bit further.

These relevant pieces of information can be in the form of a particular primary source he wants to locate, e.g. one of the 40,000+ printed documents of the Evans corpus. It can also, as is more often the case, be in the form of pieces of (background) information on the subject he is researching. This background information helps the historian to further expand the ‘information in his head’, to put the knowledge he acquired into further context, or: to develop a clearer and fuller overview and understanding of his specific research subject.

These pieces of information can vary widely, from names of authors or other agents—printers, booksellers, organizations—deemed interesting enough to check out, additional information about a document, to documents and authors similar to the ‘known entities’ already collected in the course of their research. This information-gathering goal is not necessarily a finite one, as is the case in other systems (buy book X, complete transaction Y, initiate and complete control sequence Z), but infinite, iterative.

2.3.2 The Task: identifying individual relevant items

The task the historian tries to accomplish in this iterative process can best be described as successfully identifying a potentially relevant item in a collection. The process is a cyclic, recurring one: after finishing this task, the search for other potentially relevant items may continue. It is hard to say when the overall goal is met: the individual user’s information-seeking needs can be satisfied after finding one significant piece of information. Even reaching the conclusion that there are zero items in the collection that are relevant to the user’s research can be a valid outcome of the process. It can also take numerous task cycles to meet the user's criteria, as the results of one task lead to the instigation of another task. As shown earlier, newly recognized patterns can lead to yet more new information of interest. As his contextual knowledge increases, the user develops a better understanding of the subject area. With this, the information goal is restated and refined, and can potentially lead to the formation of a new task, pointing to entirely different directions.

2.3.3 Requirements from a User Perspective

> Because information needs change in time and depend on the particular information seeker, systems should be sufficiently flexible to allow the user to adapt the information seeking process to his own current needs (Rouse and Rouse (1984), cited by Bates (1989)).

Having analyzed historians and their behavior in this chapter, leads to the following requirements for an adaptive on-line electronic finding aid, in the fashion Rouse and Rouse (1984) have in mind:

- A system for locating and accessing primary sources, not only catering to the historian's information needs and information-seeking behavior, stimulating, enabling, and empowering the researcher in this process.

- This system should be a flexible one, allowing both formal and informal search methods. It should support both (index) browsing and (keyword) searching strategies, enabling berrypicking by letting the user store the 'berries' in a 'basket'—as a digital equivalent to the index cards traditionally used. This information-collecting by picking berries enables 'internal' context-building, pattern-discovery of the 'information in his head', but can also be enhanced by making provisions for 'external' context-building and pattern-discovery—i.e. in the system—by displaying the information contextually, either textual or graphical: this is one of the potential strengths of a system's interface.

- It should also provide ample opportunity for impression formation: for the historian to get a sense of overview. This can be achieved by arranging the collection into a browsable catalog with multiple entry points, as by—personal and organizational—provenance, by subject area, by genre, by era, by location, etc, and by providing an interface and organization that is both easy to understand and helpful at the same time, and, in places highlight or provide shortcuts to important parts of the collection. Provisions should also be made to enable serendipitous discovery, for instance by providing sufficient cross-sectional links to other relevant and similar pieces of information in the collection, or in other collections.

2.4 Chapter Summary

In this past chapter we have seen that it is never safe to assume that a 'universal user' exists for such a system. There is no single Information Retrieval model that is applicable to all users. Humans sometimes behave irrationally, and do not always adhere to a strict model, however convenient it might be to assume so, especially for system developers. Humanities scholars in general, and historians in particular, behave differently than others, e.g. when seeking information for their research.

Employing a variety of informal information-seeking methods, historians display behavior which bears a close similarity to Bates' *Berrypicking* model, . In this information-seeking they primarily focus on common names in order to discover patterns and relationships in the data sources, and subsequently interesting pieces of information in a collection. These common names serve both as the nodes on which they build their 'internal' contextual framework, and as an entry point into collections. A system that sets out to cater to historians as their primary user base should be designed to accommodate, enhance and stimulate this behavior.

Chapter 3

Data: From Evans' Bibliography to Semantic Triples

In the previous chapter we have seen that the historian has a way of its own when it comes to doing research. Aside from employing different information-seeking strategies than other scholarly users of electronic finding aids, historians also seem to rely heavily on access to the primary source material for their research[1]. It is in this historical source data that they find the evidence and proof for their findings.

Thus, next to the *user* as the key ingredient[2] we can now identify a second ingredient to a successful product: the *source data*, the content. An electronic finding aid for a primary resource collection should enable the user to find and discover the historical evidence needed for his research, to pick the right 'berries'[3] from the immense 'data forest' of the collection (Bates, 1989).

In order to find good (and effective) ways to provide users access to this source data, a thorough analysis of the source data is needed. In this chapter we attempt to answer the research questions relating to the first part of the C(omputer)/system component: the data. For clarity, we will repeat these research questions here:

- What are the key characteristics of the primary source material which the finding aid to be designed centers around, the *Evans* Source?
- What is the background of the source material?
- What kind of enhancements or approaches can be derived from the nature and the organization of this material?

[1] See also sections 2.1.2 and 2.2.2. This is one of the characteristic traits of historians, albeit not exclusive: this trait is shared with other humanists as well (Stieg-Dalton and Charnigo, 2004).

[2] See section 2.1.

[3] See section 2.2.1.

The method employed in this chapter will be an analysis into both the history of *Evans* and into the historical arrangements and organization of this data source; an attempt of a bit of 'print history' of our own, as to speak. Understanding the data and the history behind it will give insight into how previous generations of historians and archivists have used this data. By looking through their eyes we can identify important (pivotal) elements in the data and grasp the taxonomical organization of the collection. This insight into the *Evans* collection will have two major uses:

1. The pivotal elements can be used to build the architecture of the finding aid's website and the underlying data structure. These elements can be used as the primary access points into the collection, as they form the main categories on which the navigational structure will be built. A finding aid which closely adheres to the way users 'consume' the data will likely be an effective one.

2. These key elements can form the basis of effective information visualizations.[4] Interactive information visualizations can play a major role in and can have a positive effect on information-seeking, context-building and cognition (Card, 2003; Shneiderman, 1996), as will be elaborated on in chapter four.

The data analysis of this chapter will look into the origin of the *Evans* collection and its inceptor, Charles Evans. We will follow the data—as well as the *meta*data[5]—from its first structured use in the *American Bibliography* through to its subsequent additions, alternative taxonomies and eventually to its inceptions crossing over into other media, like *microprint* and the *Web*. Finally, we will look into the MARC21 electronic document metadata format, which formed the basis of the latter inceptions, and provide a less document-centered alternative by restructuring the data in a Semantic triple structure.[6]

3.1 Evans, a History in Print

The story of our source data, that of the imprints within the *Evans* document set themselves, is closely linked to the story of its creator/compiler, Charles Evans.

[4] Information Visualization and its use in historical interfaces will be the subject of chapter four.

[5] According to William Arms (2000), metadata can be defined as 'data about other data, commonly divided into descriptive metadata such as bibliographic information, structural metadata about formats and structures, and administrative metadata, which is used to manage information' (Arms, 2000).

[6] Semantic triples form the underlying foundation of the SWHi Project: the Semantic Web for History.

3.1.1 Charles Evans

Born in Boston, 1850, Charles Theodore Evans devoted his life to books and libraries. He held many posts as a librarian, and helped organize many American libraries, including libraries in Boston, Baltimore, Chicago and Indianapolis, and served as the first treasurer of the American Library Association (Holley, 1963). Unfortunately, he was a hard person to work with—often in conflict with his superiors, described as hot-tempered and as a 'lone wolf'—and was often fired from his positions as a librarian (Krummel, 2005; Holley, 1963). Accordingly, in 1901, as a jobless librarian in his fifties, and with a family to provide for, Charles Evans, decided to undertake a project on a major scale: to compile a bibliography of all printed documents in Early America, from the first printing press in 1639 into the nineteenth century, to 'uncover the forgotten beginnings of the literary history of the United States' (Evans, 1935). Unsurprisingly, this massive undertaking was frowned upon by his contemporaries.

In 1903, Evans had worked his way up to the year 1729, and (privately) published his first volume of what would become known as Evans' *American Bibliography.*[7] Figure 3.1 shows the title page of this first volume.

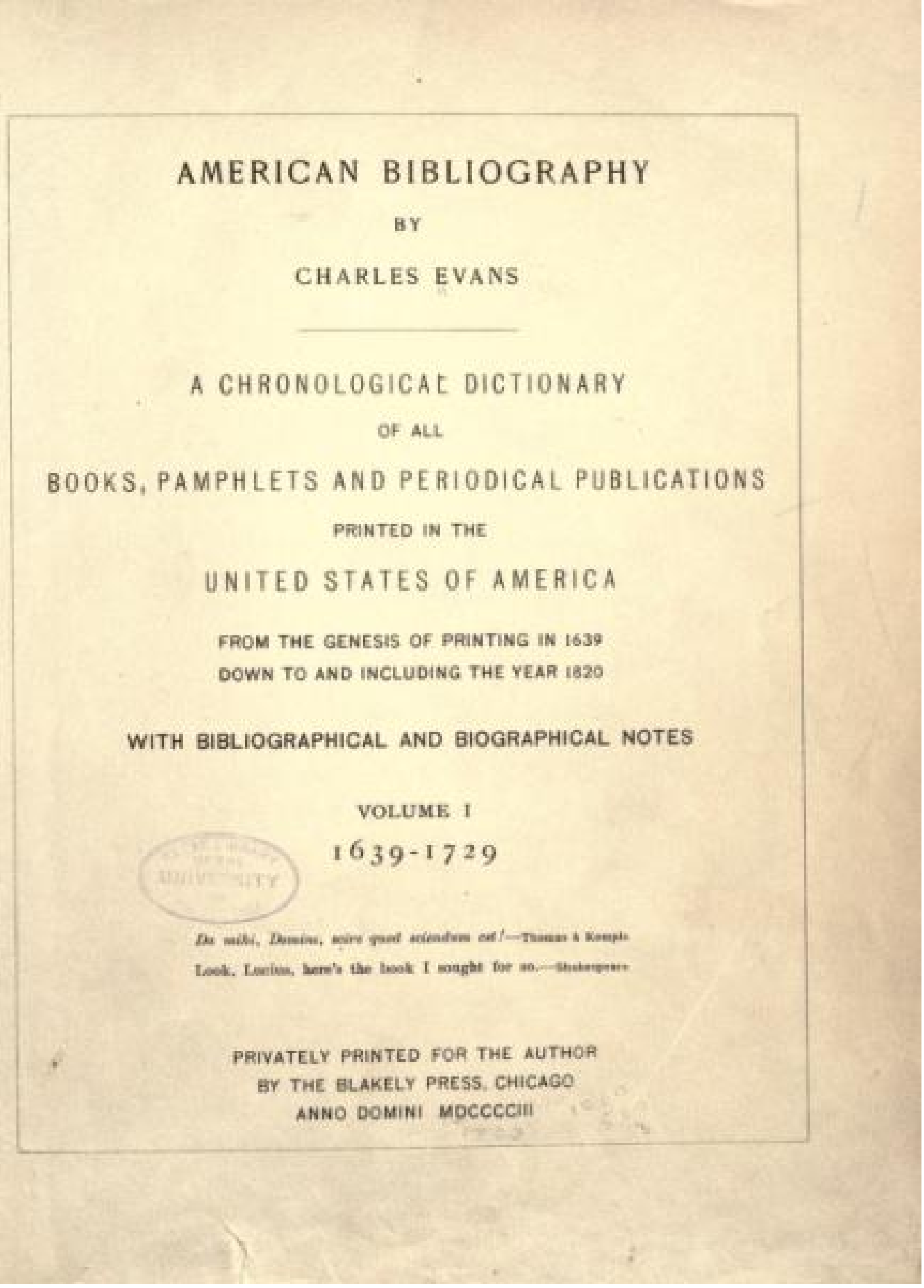
Charles Evans was not able to finish his life work. He died in 1935, only a year short of reaching the 1800 mark: his twelfth and last volume covered the years 1798 and 1799, abruptly stopping at the letter M for the year 1799. The thirteenth and final volume was completed by Clifford Shipton, and was published in 1955 by the American Antiquarian Society (AAS), which had also supported Evans during his herculean accomplishments.

In the thirteen volumes that would comprise Evans' *American Bibliography*, a total of 39,162 imprints were listed. This is a considerable achievement, even when one takes into account the fact that Evans left out certain types of print publications, including newspapers, sheet music, German Americana and some of the ephemera (Krummel, 2005).[8] An example of one of the imprints Charles Evans included into his bibliography is shown in figure 3.2.

When we take a look at the preface of the original edition of Evans' life work, his motives become clear instantly. Charles Evans wanted to document the literary history of the United States, by uncovering a 'forgotten' part of the Nation's history. He wanted to chronicle the 'birth of a National literature', by

[7] The full title of Evans' life work is *American bibliography: a chronological dictionary of all books, pamphlets, and periodical publications printed in the United States of America from the genesis of printing in 1639 down to and including the year 1820: with bibliographical and biographical notes* (Evans, 1935).

[8] According to *The Oxford Pocket Dictionary of Current English 2006*, ephemera are 'things that exist or are used or enjoyed for only a short time. Items of collectible memorabilia, typically written or printed ones, that were originally expected to have only short-term usefulness or popularity'. Clifford Shipton, in the preface of the posthumous thirteenth volume, mentions, that he, like Evans, also excluded items he considered ephemera: 'invitations, tickets, and circulars and forms which contain blank to be filled in in manuscript' (Shipton, 1955).



AMERICAN BIBLIOGRAPHY

BY

CHARLES EVANS

A CHRONOLOGICAL DICTIONARY

OF ALL

BOOKS, PAMPHLETS AND PERIODICAL PUBLICATIONS

PRINTED IN THE

UNITED STATES OF AMERICA

FROM THE GENESIS OF PRINTING IN 1639
DOWN TO AND INCLUDING THE YEAR 1820

WITH BIBLIOGRAPHICAL AND BIOGRAPHICAL NOTES

VOLUME I

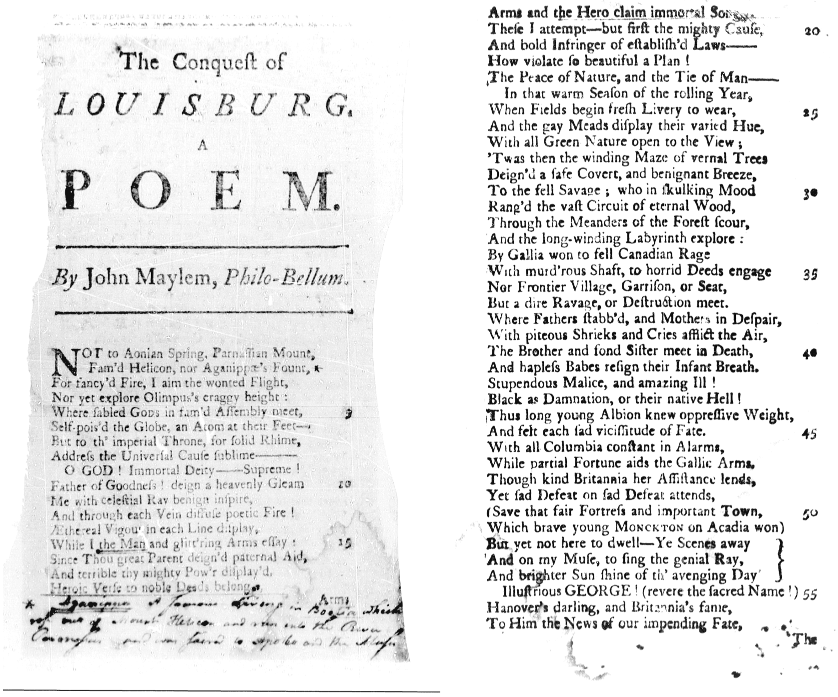
1639-1729

Da mihi, Domine, scire quod sciendum est!—Thomas à Kempis

Look, Lucius, here's the book I sought for so.—Shakespeare

PRIVATELY PRINTED FOR THE AUTHOR
BY THE BLAKELY PRESS, CHICAGO
ANNO DOMINI MDCCCCIII

Figure 3.1: Title page of the first volume of Evans' *American Bibliography.* Evans' lone undertaking was frowned upon by his contemporaries, as can be evidenced by the fact that this volume was printed privately (bottom). Source: `http://www.openlibrary.org/details/americanbibliogr01evanrich`



The Conqueſt of
LOUISBURG.
A
POEM.

By John Maylem, *Philo-Bellum.*

NOT to Aonian Spring, Parnaſſian Mount,
Fam'd Helicon, nor Aganippæ's Fount, *
For fancy'd Fire, I aim the wonted Flight,
Nor yet explore Olimpus's craggy height :
Where fabled Gods in fam'd Aſſembly meet,
Self-pois'd the Globe, an Atom at their Feet—
But to th' imperial Throne, for ſolid Rhime,
Addreſs the Univerſal Cauſe ſublime———
O GOD ! Immortal Deity——Supreme !
Father of Goodneſs ! deign a heavenly Gleam
Me with celeſtial Ray benign inſpire,
And through each Vein diffuſe poetic Fire !
Æthereal Vigour in each Line diſplay,
While I the Man and glitt'ring Arms eſſay :
Since Thou great Parent deign'd paternal Aid,
And terrible thy mighty Pow'r diſplay'd,
Heroic Verſe to noble Deeds belong.
Arms

* Aganippe a famous Spring in Boeotia which [illegible] out of Mount Helicon and ran into the River [illegible] and was sacred to Apollo and the Muses

Arms and the Hero claim immortal Song.
Theſe I attempt—but firſt the mighty Cauſe,
And bold Infringer of eſtabliſh'd Laws——
How violate ſo beautiful a Plan !
The Peace of Nature, and the Tie of Man——
In that warm Seaſon of the rolling Year,
When Fields begin freſh Livery to wear,
And the gay Meads diſplay their varied Hue,
With all Green Nature open to the View ;
'Twas then the winding Maze of vernal Trees
Deign'd a ſafe Covert, and benignant Breeze,
To the fell Savage ; who in ſkulking Mood
Rang'd the vaſt Circuit of eternal Wood,
Through the Meanders of the Foreſt ſcour,
And the long-winding Labyrinth explore :
By Gallia won to fell Canadian Rage
With murd'rous Shaft, to horrid Deeds engage
Nor Frontier Village, Garriſon, or Seat,
But a dire Ravage, or Deſtruction meet.
Where Fathers ſtabb'd, and Mothers in Deſpair,
With piteous Shrieks and Cries afflict the Air,
The Brother and fond Siſter meet in Death,
And hapleſs Babes reſign their Infant Breath.
Stupendous Malice, and amazing Ill !
Black as Damnation, or their native Hell !
Thus long young Albion knew oppreſſive Weight,
And felt each ſad viciſſitude of Fate.
With all Columbia conſtant in Alarms,
While partial Fortune aids the Gallic Arms,
Though kind Britannia her Aſſiſtance lends,
Yet ſad Defeat on ſad Defeat attends,
(Save that fair Fortreſs and important Town,
Which brave young Monckton on Acadia won)
But yet not here to dwell—Ye Scenes away
And on my Muſe, to ſing the genial Ray,
And brighter Sun ſhine of th' avenging Day
Illuſtrious GEORGE ! (revere the ſacred Name !)
Hanover's darling, and Britannia's fame,
To Him the News of our impending Fate,
The

Figure 3.2: Sample pages for *The conquest of Louisburg*, by John Maylem (1758) (Imprint no. 14254). For purposes of clarity, we shall use the same example throughout this chapter, the imprint titled *The conquest of Louisburg*, by John Maylem (1758; Evans item no. 14254). Source: `http://opac.newsbank.com/select/evans/14254`

shining a light on a period 'all to little studied' (Evans, 1935).[9] In this, Evans is a child of his age, in which the rise of nation states, national awareness and cultural nation-thinking brought with it a surge in interest to uncover, retrace and document the early beginnings of the Nation. But that in itself can be the subject of an interesting—but separate—study. Our focus should remain on Evans' *American Bibliography*, especially on the works contained within it.

3.1.2 Evans' American Bibliography

When compiling a bibliography in print, a major consideration for a bibliographer is choosing what variable to use to arrange the bibliographical listings. The print format is fairly inflexible in nature, and as space is very much limited, there is

[9] Almost all instances in which we quote Charles Evans were taken from the preface to his first volume (1903). Charles Evans included a lengthy preface to each volume, which are interesting reading material. When read back to back, all prefaces combined, provide an interesting documentation to Evans' process of compiling his life work, giving a unique insight into his considerations. They also provide an excellent overview of the print history of each time period, as written and highlighted by Charles Evans himself.

usually only room for one primary arranging order, and the annotations for each imprint should be kept brief. This is especially the case with a bibliography that tries to document all known imprints of an entire continent, from a collection spanning the better part of two centuries, a collection comprising tens of thousands of documents. Charles Evans arranged his bibliography—as its subtitle suggests[10]—primarily in a chronological way:

> the date of publication is the most important fact in the identification of books and editions; it is the key to all investigation. With its aid it is possible to unlock any clew of publication, and even the mutilated fragment yields its secret to the trained investigator. [...] For remembrance, we repeat, the fact first in importance in bibliographical research is the date—always the date! (Evans, 1935).

From this we can infer Evans' pivotal document characteristic, or—in terms of an (electronic) finding aid—the primary access point: the imprints are organized by year. Here we see an early example that the goals of a bibliographer and a historian do not always fully align: whereas the historians are interested in the contents of an imprint, and seem to prefer to access information by *provenance*[11], the bibliographer has other (organizational) considerations—and other potential audiences—to take into account, and has—in this case—chosen for a chronological arrangement. To a bibliographer, other properties of an imprint are of interest: 'The format, binding, paper, types, the author, title, printer, date, to him are interesting traits of book character' (Evans, 1935).

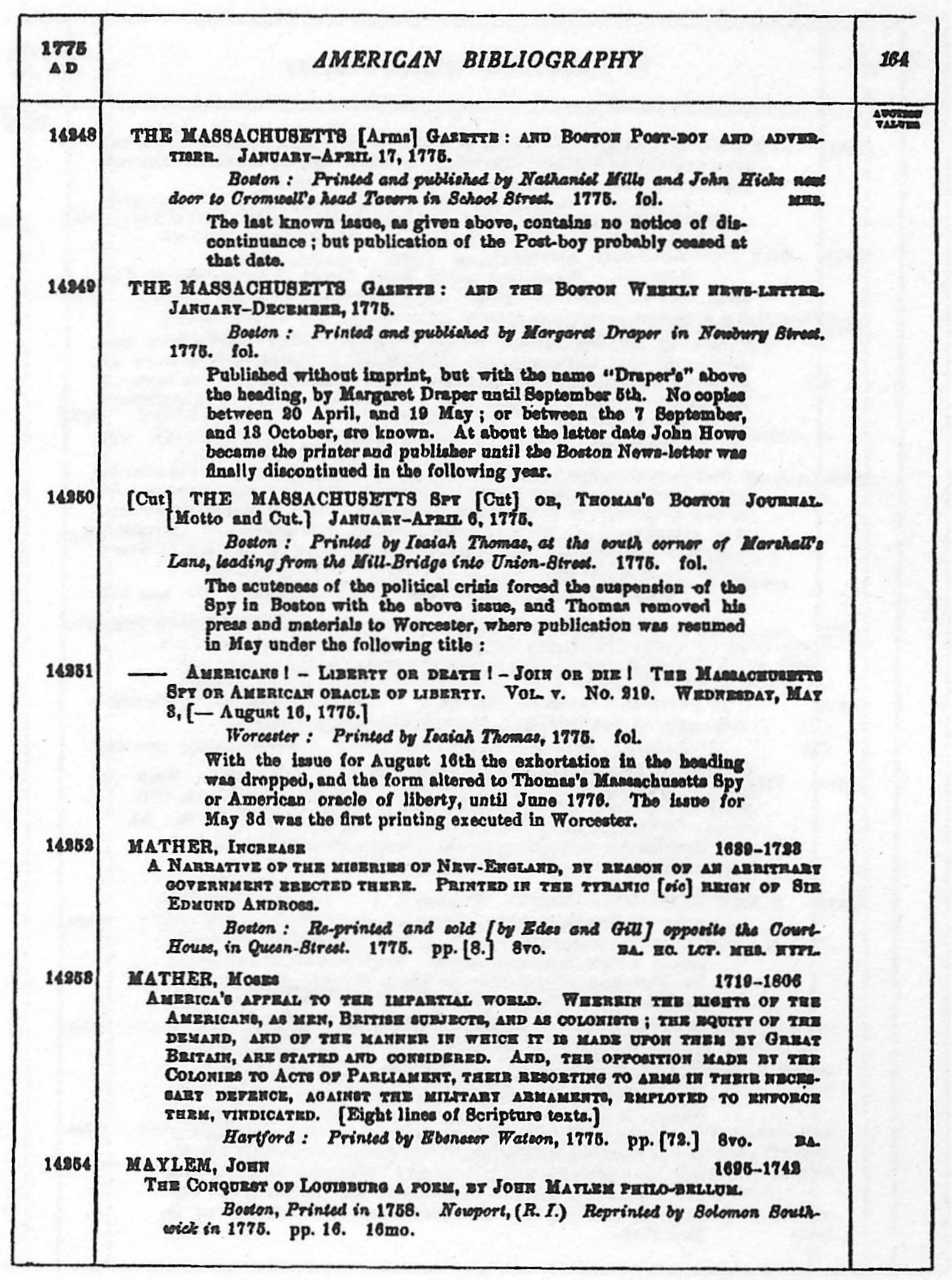
However, scholars in search of documents relating to particular people or organizations can also find their way through the collection fairly easily: within each year of publication, by name (provenance): 'under the author if known, under its title if anonymous' (Evans, 1935).[12] Within this listing for each author for that particular year, the imprints are arranged alphabetically, by title. An example of this arrangement is shown in figure 3.3.

Figure 3.3 shows brief entries for each author and for each imprint. The most prominently featured parts for each imprint are the year of the imprint's publication, its author, and the imprint's title. These three properties of each imprint are used by Evans to arrange the bibliography, and can therefore be considered the

[10] Charles Evans' *American Bibliography* is subtitled 'a chronological dictionary of all books, pamphlets, and periodical publications printed in the United States of America from the genesis of printing in 1639 down to and including the year 1820 : with bibliographical and biographical notes'.

[11] See also section 2.2.2. on name-collecting and provenance.

[12] 'Author' in this context should be seen in a large sense. It can refer to both individuals and organizations. Access to a printing press was—especially in the early years—for the most part was the exclusive domain of a small number of organizations (mainly government, university, or religious). Also, most individuals, as authors of these early American imprints, can be linked to such an organization.



1775 A D | *AMERICAN BIBLIOGRAPHY* | 164

Auction values

14248 THE MASSACHUSETTS [Arms] Gazette : and Boston Post-boy and advertiser. January–April 17, 1775.

Boston : Printed and published by Nathaniel Mills and John Hicks next door to Cromwell's head Tavern in School Street. 1775. fol. mhs.

The last known issue, as given above, contains no notice of discontinuance ; but publication of the Post-boy probably ceased at that date.

14249 THE MASSACHUSETTS Gazette : and the Boston Weekly news-letter. January–December, 1775.

Boston : Printed and published by Margaret Draper in Newbury Street. 1775. fol.

Published without imprint, but with the name "Draper's" above the heading, by Margaret Draper until September 5th. No copies between 20 April, and 19 May ; or between the 7 September, and 13 October, are known. At about the latter date John Howe became the printer and publisher until the Boston News-letter was finally discontinued in the following year.

14250 [Cut] THE MASSACHUSETTS Spy [Cut] or, Thomas's Boston Journal. [Motto and Cut.] January–April 6, 1775.

Boston : Printed by Isaiah Thomas, at the south corner of Marshall's Lane, leading from the Mill-Bridge into Union-Street. 1775. fol.

The acuteness of the political crisis forced the suspension of the Spy in Boston with the above issue, and Thomas removed his press and materials to Worcester, where publication was resumed in May under the following title :

14251 —— Americans ! – Liberty or death ! – Join or die ! The Massachusetts Spy or American oracle of liberty. Vol. v. No. 219. Wednesday, May 3, [— August 16, 1775.]

Worcester : Printed by Isaiah Thomas, 1775. fol.

With the issue for August 16th the exhortation in the heading was dropped, and the form altered to Thomas's Massachusetts Spy or American oracle of liberty, until June 1776. The issue for May 3d was the first printing executed in Worcester.

14252 MATHER, Increase 1639–1723

A Narrative of the miseries of New-England, by reason of an arbitrary government erected there. Printed in the tyranic [*sic*] reign of Sir Edmund Andross.

Boston : Re-printed and sold [by Edes and Gill] opposite the Court-House, in Queen-Street. 1775. pp. [8.] 8vo. ba. hc. lcp. mhs. nypl.

14253 MATHER, Moses 1719–1806

America's appeal to the impartial world. Wherein the rights of the Americans, as men, British subjects, and as colonists ; the equity of the demand, and of the manner in which it is made upon them by Great Britain, are stated and considered. And, the opposition made by the Colonies to Acts of Parliament, their resorting to arms in their necessary defence, against the military armaments, employed to enforce them, vindicated. [Eight lines of Scripture texts.]

Hartford : Printed by Ebenezer Watson, 1775. pp. [72.] 8vo. ba.

14254 MAYLEM, John 1695–1742

The Conquest of Louisburg a poem, by John Maylem philo-bellum.

Boston, Printed in 1758. *Newport, (R. I.) Reprinted by Solomon Southwick in* 1775. pp. 16. 16mo.

Figure 3.3: Sample page of Evans' *American Bibliography*, showing Evans' short descriptions of each imprint. On the bottom of the page is imprint no. 14254, Maylem's *Conquest of Louisburg.* Source: Evans (1935)

primary *metadata* variables—according to Charles Evans at least—of the *Evans* collection. Furthermore, each separate imprint was assigned its own unique serial number, for 'convenience of reference', without 'the inconvenience of long titular reference' (Evans, 1935).

Along with these primary document properties, Charles Evans chose to include the imprint's publication information (printer, location, number of pages, format). If the document's author was known, Evans also included the person's years of birth and death. Some entries appear with annotations, included by Charles Evans to clarify possible issues with different editions, or the continuity of a particular periodical imprint.

The main organization of the entries by Evans in his *American Bibliography* can be seen, in terms of information-seeking, as an 'access point'. The pivotal metadata variable *year* can be seen as the primary access point of the *American Bibliography.*

However, Charles Evans does provide us with other points of entry into the collection, including the provenance-based access point favored by historians. At the end of each volume Charles Evans included two auxiliary indices, or 'secondary' access points. The first—and most voluminous—secondary access point is what Evans calls a 'briefer *Index of Authors*'. This arrangement based on provenance—on common names—has 'the duty of keeping the works of each author together', which is of minor importance than arrangement by chronology. It only has one purpose—according to Evans—which is the 'sole advantage of an alphabetical arrangement'(Evans, 1935). The other secondary access point Charles Evans deemed useful is a classified *subject index*, illustrating 'the wide range of interests which engaged the Fathers of the Republic'. Each entry in both indices includes a reference to the 'Evans numbers' of the corresponding imprints.

Along with these two auxiliary indices one other list emerges from the back pages as a potential source of interest to historians: Charles Evans deemed it of 'bibliographical interest and importance' to include a *list of printers and publishers*, arranged per colony/state/province, and per city. In this list Evans did not include cross-references to the imprints themselves, but it does serve another very useful purpose to (print) historians: it documents the spread of printing in North America. For every city and state, along with the list of printers, a year is included. This year, although seemingly insignificant at first, denotes when printing was instituted in a particular city and state (for example 'Printing instituted in Massachusetts, in 1639') (Evans, 1935).

Charles Evans' index of printers shows that first printing press made its way into British North America a mere 19 years after the Pilgrims first set foot in the New World.[13] The first printing press was located in the present-day state

[13] On a historical side note: The Pilgrims built their first settlement in present-day Massachusetts, and called it Plymouth. The area was inhabited by several native Algonquian tribes. The Pilgrims were followed by—vehemently religious—Puritan settlers who established the Massachusetts Bay Colony, near present-day Boston. In this colony, the puritan leader and

of Massachusetts (1639, in Cambridge, followed by Boston (1675), and Plymouth and Charlestown (1685)). It took over four decades for printing presses to arrive in other states, when first Pennsylvania (1685) and eight years later New York (1693) followed suit.[14]

This might seem irrelevant data to the unsuspecting reader, but to a historian interested in Early American imprints this is really important and potentially insightful information. Evans himself certainly deemed it interesting enough to include, and it is information like this that can give us insight into what is important in the *Evans* collection.

3.1.3 Additions to the *American Bibliography*

As mentioned earlier, Charles Evans passed away before being able to complete all volumes of his life work, and before finishing it off with one full index encompassing the entire corpus. The original volumes of *Evans* contained 39,162 titles, but there was no guide yet to the 10,035 additional titles that were uncovered, by Roger Bristol and by the AAS, making up a total of 49,197 entries (Bristol, 1970; Shipton and Mooney, 1969).

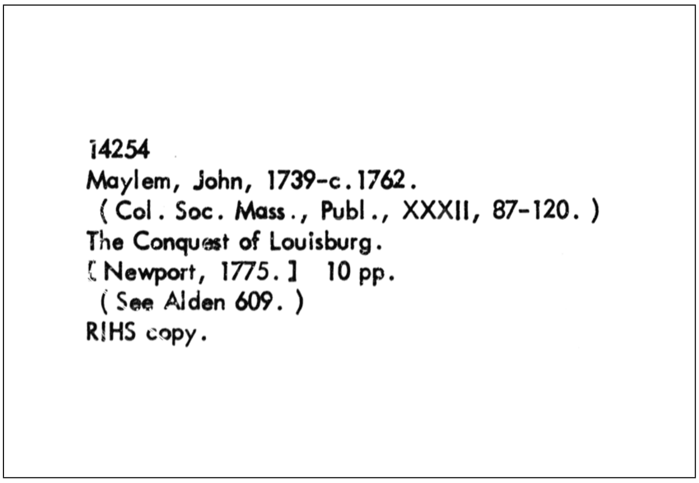
In 1954, when the idea was coined to reproduce in microprint the full text of every non-serial item listed by Evans—along with Bristol's supplement—, Clifford Shipton—who had completed the thirteenth volume of Evans's life work—and James Mooney of the AAS took the work upon them to compile the full index Charles Evans had aspired to make. This *Short-title Evans*, as it would be titled, would simultaneously serve as the finding aid to accompany the microprint edition. In the process of compiling the *Short-title Evans*, Shipton and Mooney made tens of thousands of bibliographical corrections, mostly having to do with authorship attribution, anonymous items, and 'ghost items'. Charles Evans, who did not have the physical imprints themselves to work with, made a lot of assumptions, on unbroken chronological runs of almanacs and government publications for instance, and on booksellers advertisements implying separate editions of imprints. For the original 39,162 *Evans* items, Shipton and Mooney (1969) concluded that about one in ten was a ghost or contained a serious bibliographical error.[15]

colony governor John Winthrop founded the first university on North-American soil, Harvard University, in 1639. It is interesting to compare this historical information with one of our visualizations, the 'Zeitgeist' subject cloud for the first full decade of printing, the 1640's, similar to the one shown in figure 3.9, available on `http://evans.ub.rug.nl/~peter/test/zeitgeist.php?time=1640&type=decade`.

[14] For those interested at the early history of American printing, William S. Reese's *The First Hundred Years of Printing in British North America: Printers and Collectors* (Reese, 1990), `http://www.reeseco.com/papers/first100.htm` is an interesting read.

[15] This accounts for the major discrepancy between the total number of items in *Evans*, and that of the data set we are working with, which contains a 'mere' 35,823 imprint instances.

Shipton and Mooney set out to create a collection of *index cards*: one for every imprint, *and* one for every author (See figure 3.4). They arranged these tens of thousands of index cards alphabetically in the resulting *Short-title Evans.* In order to create a less voluminous bibliography, they shortened the bibliographic entries: the titles were shortened, and they added no extra descriptions, other than where they wanted to indicate a difference with Evans's bibliographical description.



14254
Maylem, John, 1739-c.1762.
(Col. Soc. Mass., Publ., XXXII, 87-120.)
The Conquest of Louisburg.
[Newport, 1775.] 10 pp.
(See Alden 609.)
RIHS copy.

Figure 3.4: Index card for Evans imprint no. 14254, showing the basic metadata used by Shipton and Mooney (1969). Source: `http://opac.newsbank.com/select/evans/14254`

One other bibliographical collection is important to mention here: As one might recall, Charles Evans' original aim was to complete the listing of imprints all the way to the year 1820. The thirteenth volume of Evans, as compiled by Roger Bristol, reached the year 1800. The imprints for the remaining years were compiled by Ralph Shaw and Richard Shoemaker. The *Shaw-Shoemaker* corpus is universally regarded as the twin companion to *Evans*.

3.1.4 Alternative Indices to *Evans*

Over the years, *Evans*, Bristol's *Supplement*, and Shipton and Mooney's *Short-title Evans* were joined by a number of special-interest indices of early American imprints, like Bristol's *Index of Printers and Booksellers*, and Tanselle's *Guide* on the same topic, several indices of Maps (both by the AAS itself and by Jim Walsh). Maps are interesting entities in a historical imprint collection: sometimes, they are indexed by themselves, as Atlases, but more often they are contained within an imprint, for instance when a map is used as an illustration in a history book or in a government publication.

Other bibliographies and indexes—partially or fully based on the *Evans* corpus—have been published on several specialized subjects, like medical publications, textiles, and on imprints in specific languages—like German and Dutch—and imprints concerning specific geographic entities—like the one for imprints regarding Rhode Island. The existence of these specialized, single-topic, bibliographies indicate that there is a need for highlighting special or interesting items in the collection, to accommodate for the varying preferences of the different user groups of the *Evans* collection.

3.1.5 *Evans*, the Next Generation: Digital Editions

With the inception of the Microprint edition (an effort by the AAS, lead by Clifford Shipton) in 1954, the bibliography compiled by Charles Evans became a collection, in its 'proper' sense: the documents themselves became available in a single collection. Researchers around the world—whose access to the imprints was previously restricted to the brief descriptions provided in the *American Bibliography*—could, for the first time, gain access to the source documents themselves, from the comfort of their own university library. The various additions and alternative indices made it possible to find the information needed relatively easily. But still, ready access was still restricted to a physical location: a (university) library with a subscription and a copy of the Microprint edition. Furthermore, quick and ready access to the imprints needed still remained an elaborate, time-consuming process. Among other things, it depended on which finding aid—*American Bibliography*, *Short-title Evans* and/or other—was available, and 'stumbling upon' potential items of interest depended on 'flipping through' a large amount of slides and pages, as well as on sheer luck. With the advent of personal computers, however, and of the emergence of electronic finding aids, new possibilities arose. The digital age brought the information into the peoples' homes and onto the scholars' desks by way of the personal computer and the World Wide Web. A digital (on-line) version of (historical) document collection—combining the collection and the finding aid into one—has the potential of unlimited flexibility and personalization, and can be accessed from anywhere.

Evans Digital Edition

The *Early American Imprints, Series I: Evans (1639-1800)*[16], or *Evans Digital Edition*, as it is also known, is the next generation of the *Evans* collection and its indices and finding aids. It is published and maintained by Readex, a division of the Newsbank corporation which specializes in maintaining document collections, both in digital and in microprint formats. It is available by subscription only, and is a part of the larger *Archive of Americana*, which also includes *Early American Imprints, Series II: Shaw-Shoemaker, 1801-1819*, a continuation and

[16] The *Early American Imprints, Series I: Evans (1639-1800)* collection is available on the Web at `http://infoweb.newsbank.com/?db=EVAN&s_start=evans`, and is available through (library) subscription only.

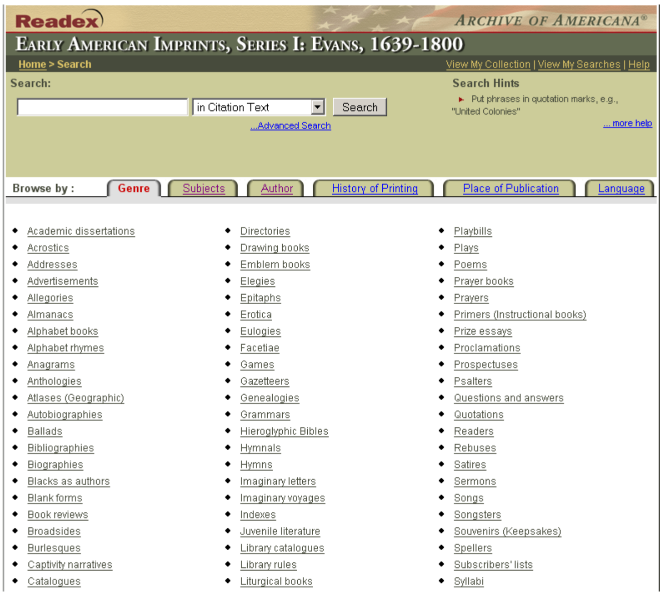
near completion of what Charles Evans originally set out to accomplish: to try to reach the year 1820. Readex' *Archive of Americana* also encompasses collections of the main document types originally left out of *Evans* by its inceptor, namely newspapers and ephemera.

The web-based collection of the *Evans Digital Edition* supports information-seeking through both keyword searching and browsing. Browsing is accommodated by implementing a variation of the 'directory' or 'catalog' model (See figure 3.5).[17] The *Digital Edition*'s directory provides a classification into six main categories: *genres*, *subjects*, *author*, *history of printing* (printer/publisher/bookseller), *place of publication*, and *language*. All six come with a large number of subcategories to choose from. The *Digital Edition*'s main strength, as compared to earlier inceptions of *Evans*, is that a web-based interface can combine both the finding aid as the digital facsimiles and full texts of the historical documents themselves. At the same time, a web-based interface can be made available to researchers all over the globe—provided they have a subscription, of course.

A number of observations can be made as to the organization of the categories, as well as to the consistency of the digital collection's primary access points. This might not seem entirely relevant in the analysis of data and of the collection, but it might indicate certain underlying assumptions made by the digital collection's developers on which categories and ordering arrangements they consider most important. For example, the first page, the primary access point of the collection is generally an overview page of the collection. Usually the 'home' page of a collection is the 'top node' in the hierarchical browsing tree of the collection, providing an overview of the most important entry points—usually represented by categories—of a collection. Such an overview page (sometimes called 'index' or 'portal' page) is not present in the *Evans Digital Edition*, however. Instead the visitor enters the *Evans* collection via the Genres category overview page, one level below the top node (See figure 3.5). This implies that the creators of the *Digital Edition* rank the *genres* categorization, not *author/name* (as suggested by our analysis of the historians' information-seeking behavior), or *year* (as suggested by Evans), as the most important access point. These are, however, available as secondary access points, as are other ordering arrangements and categorizations. They remain accessible on demand, by clicking on one of the other categories shortcut links on the collection's web page.

The flexibility of an electronic version of a document collection and document finding aid, allows for a nearly endless availability in categories, and of subdivisions thereof. It allows for a wider range of options to present and arrange and order the collection data, and is in no way subject to the limitations of the bibliography in print, which forced Charles Evans and his colleagues to

[17] The 'web directory' model was popularized by one of the most popular search engines of the mid-1990s, Yahoo! (`http://dir.yahoo.com/`). Other current examples are the Google Directory (`http://www.google.com/Top/`) and the Open Directory project (`http://dmoz.org/`).



- Academic dissertations
- Acrostics
- Addresses
- Advertisements
- Allegories
- Almanacs
- Alphabet books
- Alphabet rhymes
- Anagrams
- Anthologies
- Atlases (Geographic)
- Autobiographies
- Ballads
- Bibliographies
- Biographies
- Blacks as authors
- Blank forms
- Book reviews
- Broadsides
- Burlesques
- Captivity narratives
- Catalogues
- Directories
- Drawing books
- Emblem books
- Elegies
- Epitaphs
- Erotica
- Eulogies
- Facetiae
- Games
- Gazetteers
- Genealogies
- Grammars
- Hieroglyphic Bibles
- Hymnals
- Hymns
- Imaginary letters
- Imaginary voyages
- Indexes
- Juvenile literature
- Library catalogues
- Library rules
- Liturgical books
- Playbills
- Plays
- Poems
- Prayer books
- Prayers
- Primers (Instructional books)
- Prize essays
- Proclamations
- Prospectuses
- Psalters
- Questions and answers
- Quotations
- Readers
- Rebuses
- Satires
- Sermons
- Songs
- Songsters
- Souvenirs (Keepsakes)
- Spellers
- Subscribers' lists
- Syllabi

Figure 3.5: The Evans Digital Edition entry page. The entry page leads the visitor directly into the *Genre*-category. Source: `http://infoweb.newsbank.com/?db=EVAN`

make stringent—sometimes arbitrary—(design) decisions on what information to include and how to order it. An analysis of the *Evans Digital Edition* is not complete without looking at the design decisions made by its developers, and of their choices in how to subdivide and arrange the *Evans* imprint collection.

The *Genre* category, which is most prominently featured in the *Digital Edition*, lists all imprints according to genre or type of publication. 'Genre' sometimes denotes print formats (Broadsides), but in general it refers to the type of content of an imprint (Academic Dissertations, Poems, Children's literature, Cookbooks, Psalms) Every genre has its own separate subcategory listing; there are 91 genres in total. Imprints can belong to multiple genre subcategories. Within each genre subcategory, imprints are listed chronologically.

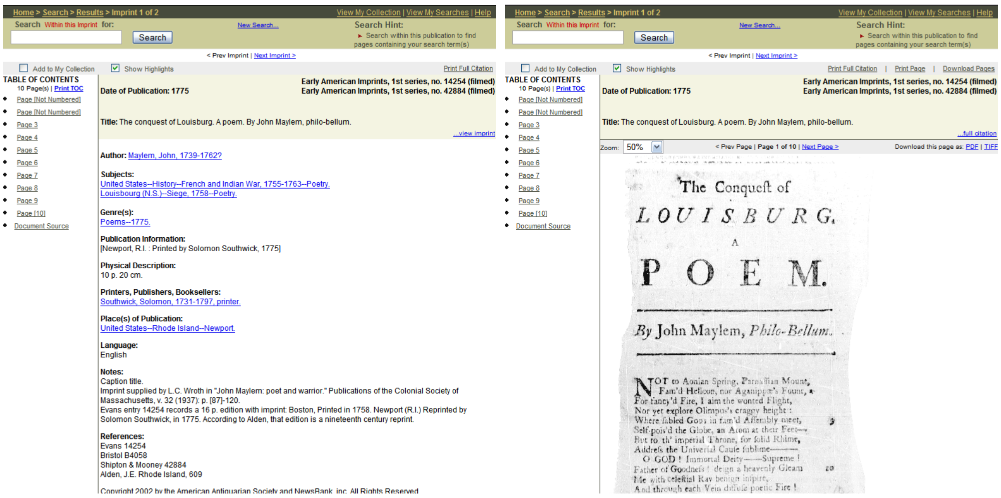
Subjects are divided into sixteen subcategories: Economics and Trade, Government, Health, History, Labor, Languages, Law and Crime, Literature, Military, Peoples, Philosophy, Politics, Religion, Science, Society, Manners and Customs, and Theology. Each of the sixteen subject subcategories has a large number of subcategories itself, ranging from a few dozen to well over several hundreds of subjects per subcategory. For example, the subject category 'Economics and Trade' has 224 'sub-subjects', in topics ranging from Tea Tax and Steamboats to Coinage and Contraband of War.

The *Author* category is organized by (last) name, and is subdivided into twenty-six subcategories: one for each letter of the alphabet. Each subcategory page which is divided into three subsections: Conferences[18], Organizations,

[18] In this context, a conference denotes a (temporary) convention of people, like the Assembly of Pastors of Churches in New-England of 1743, the annual Convention of Delegates from the

and People. The category *History of Printing* is organized in a similar fashion: the three subdivisions in this category are those of Printers, Publishers, and Booksellers. *Place of Publication* has a more hierarchical three-tiered structure: Country, State (Province), and City. In the *Language* category we find all the languages present in *Evans*, ordered alphabetically. [19]

The focus of the entire *Digital Edition* is solely on documents, on imprints. Although there are several categories devoted to people (authors, printers, booksellers) and places (location), there are no detail pages for these entities. On the lowest node of each category, when there are no further subcategories available, the only thing that can be found is a chronological listing of all imprints belonging to that subcategory. The same is the case for search results pages. A search for 'Washington' does not return any person (George, Martha Washington) or location matches (Washington D.C., Washington state), only document matches. The search results are also ordered chronologically, not by 'relevance' as is the case in most search engines based on the traditional IR model.[20] No alternative sorting options (i.e. alphabetic, popularity (most viewed), relevance) are provided.



Early American Imprints, 1st series, no. 14254 (filmed)
Early American Imprints, 1st series, no. 42884 (filmed)

Date of Publication: 1775

Title: The conquest of Louisburg. A poem. By John Maylem, philo-bellum.

Author: Maylem, John, 1739-1762?

Subjects:
United States--History--French and Indian War, 1755-1763--Poetry.
Louisbourg (N.S.)--Siege, 1758--Poetry.

Genre(s):
Poems--1775.

Publication Information:
[Newport, R.I. : Printed by Solomon Southwick, 1775]

Physical Description:
10 p. 20 cm.

Printers, Publishers, Booksellers:
Southwick, Solomon, 1731-1797, printer.

Place(s) of Publication:
United States--Rhode Island--Newport.

Language:
English

Notes:
Caption title.
Imprint supplied by L.C. Wroth in "John Maylem: poet and warrior." Publications of the Colonial Society of Massachusetts, v. 32 (1937): p. [87]-120.
Evans entry 14254 records a 16 p. edition with imprint: Boston, Printed in 1758. Newport (R.I.) Reprinted by Solomon Southwick, in 1775. According to Alden, that edition is a nineteenth century reprint.

References:
Evans 14254
Bristol B4058
Shipton & Mooney 42884
Alden, J.E. Rhode Island, 609

Copyright 2002 by the American Antiquarian Society and NewsBank, inc. All Rights Reserved

The Conqueſt of
LOUISBURG.
A
POEM.
By John Maylem, *Philo-Bellum.*

NOT to Aonian Spring, Parnaſſian Mount,
Fam'd Helicon, nor Aganippe's Fount,
For fancy'd Fire, I aim the wonted Flight,
Nor yet explore Olimpus's craggy height:
Where fabled Gods in fam'd Aſſembly meet,
Self-pois'd the Globe, an Atom at their Feet—
But to th' imperial Throne, for ſolid Rhime,
Addreſs the Univerſal Cauſe ſublime———
O GOD! Immortal Deity——Supreme!
Father of Goodneſs! deign a heavenly Gleam 10
Me with celeſtial Ray benign inſpire,
And through each Vein diffuſe poetic Fire!

Figure 3.6: *Evans Digital Edition* detail pages for Imprint no. 14254 (left). The hyperlinks on the left side of the page provide access to the scanned pages of the imprint (right). Source: `http://opac.newsbank.com/select/evans/14254`

Keyword searching is possible in two modes: within the entire collection, and within a subset of the collection (an earlier search result set, a subcategory, within a single document). Keyword searching is not restricted to the basic metadata;

Abolition Societies, and the 1662 Boston Synod.

[19] There are eleven languages present in *Evans*, according to the *Digital Edition* overview page: Algonquian, Dutch, English, French, German, Greek(Ancient, to 1453), Latin, Mohawk, Spanish, Swedish, and Welsh.

[20] See section 2.2.1 on Information Retrieval.

it can also be performed within the entire full-text corpus of the collection, which is available for the majority of *Evans* imprints.[21] On the detail pages all imprint metadata is provided, as well as crosslink shortcuts to the subcategory pages of the most important metadata elements (See figure 3.6, left). The *Evans Digital Edition* also allows its users to click through all pages of the (scanned) imprint, as well as the option to download them, as is shown in figure 3.6, on the right.

Krummel's taxonomy

An alternative taxonomy for an on-line electronic finding aid for the *Evans* collection was proposed in 2005, by Donald Krummel (Krummel, 2005). Coming from a printing history background, Krummel proposed 'an on-line system based based on Tanselle's[22] contents page'. Krummel proposed—based on one of the alternative overviews of early American imprints, Tanselle's *Guide to the study of United States imprints*—to create an on-line index based on several kinds of access points, 'accessible through a menu of fearsome complexity'.

The access points proposed by Krummel (2005)—at least the ones relevant to the design of an on-line finding aid—are:

1. names of printers and publishers;
2. geographical dimensions, in which 'cities also need to be accessible by state and region';
3. imprint dates, accessible by era;
4. names of authors,
5. (names of) genres, and
6. (names of) subjects.

Krummel also stressed that 'considering today's interest in the history of reading and ownership, references should be made to the evidence when it exists', i.e. linking to the (facsimiles of the) original documents. (Krummel, 2005).

Most of the *Digital Edition*'s categories can also be found in Krummel's taxonomy. The main difference between this taxonomy and the directory model employed by the current *Evans Digital Edition* lies mostly in its emphasis on

[21] The Full-text *Evans* corpus is provided by a collaborative effort of a large number of scholarly institutions called the Text Creation Partnership(`http://www.lib.umich.edu/tcp/evans/`). The TCP has—in collaboration with Newsbank and Readex—encoded most of the full texts of *Evans* into a searchable digital text format.

[22] Krummel is referring to the Tanselle *Guide*: Tanselle (1971), *Guide to the study of United States imprints*, the '*vade mecum* for all those interested in the history of American printing and publishing'. (Riley, 1971).

the provenance and the print history of the imprint: *When* and *where* was this edition made, and *by whom*? Krummel's taxonomy adds one important entry point to the proposed directory structure: *imprint dates, accessible by era.*[23]

3.2 Transforming Metadata

As evidenced by the previous sections, bibliographic finding aids (both in paper and electronic formats) are powered by metadata. As mentioned earlier, metadata is data about data. On the Web, metadata is used to describe information about a specific website—information that is not necessarily on the web page itself—like the author, publication date, language, or a short description and a brief keyword summary. This information is used by browsers, search engines, and other websites to interpret the web page itself, for indexing or display purposes. Books and other bibliographic items also have metadata. An imprint has an author, a publisher, an ISBN, and a price; it also has a title, a subtitle, a specific publication format, and a number of pages. When buying a book from an on-line book store, or when lending a book from a library, the metadata will be the information one will use to locate the book.

3.2.1 Imprint Metadata: MARC21

In the 1960s and 1970s various efforts were undertaken to develop a format to describe bibliographic metadata, for use in electronic (computerized) library catalogs. These efforts were undertaken mainly by national libraries, and as a result each country ended up with its own format, all with naming variations on 'MARC': the American Library of Congress developed USMARC, Canada developed CAN/MARC, and various European countries also had their own variety of MARC. The acronym MARC stands for MAchine-Readable Cataloging, and defines a bibliographic data format. In 1997 the American and Canadian MARC offices decided to combine and harmonize their MARC variations, to develop a standard format, which they named MARC21. This MARC21 format has since become the de facto standard.[24] The MARC21 format provides the mechanism by which computers exchange, use, and interpret bibliographic information.[25]

A MARC21 record contains all metadata needed by libraries to describe bibliographic material. It is structured in accordance with the international ISO

[23] Other logical temporal subdivisions are an event-based taxonomy or a hierarchical one (century, decade, year).

[24] Many countries still use their own varieties of MARC. Many—but not all—are either a translation or an adaptation of MARC21. Marc Standards office, 'MARC Translations', `http://www.loc.gov/marc/translations.html`.

[25] This information was adapted from the Official MARC Standards website, maintained by the Library of Congress (`http://www.loc.gov/marc/`).

2709[26], and its contents are made up of several coded variable fields and subfields. Listing 3.1 shows part of a sample MARC21 record, for *Evans* imprint no. 14254 (John Maylem, *The conquest of Louisburg* (1775)), encoded in MarcXML format.[27]

Listing 3.1: Sample MARC21 Record (part) for Imprint no. 14254 (MarcXML format)

```
<record type="Bibliographic">
<leader>02263cam  22003971a 4500      </leader>
[...]
<controlfield tag="008">850520s1775    riu     s      000 0 eng d          </controlfield>
[...]
<datafield tag="100" ind1="1" ind2="">
<subfield code="a">Maylem , John ,</subfield>
<subfield code="d">1739-1762?      </subfield>
</datafield>
<datafield tag="245" ind1="1" ind2="4">
<subfield code="a">The conquest of Louisburg .</subfield>
<subfield code="h">[electronic resource] :</subfield>
<subfield code="b">A poem . /</subfield>
<subfield code="c">By John Maylem , philo-bellum .                          </subfield>
</datafield>
<datafield tag="260" ind1="" ind2="">
<subfield code="a">[Newport , R.I. :</subfield>
<subfield code="b">Printed by Solomon Southwick ,</subfield>
<subfield code="c">1775]            </subfield>
</datafield>
<datafield tag="300" ind1="" ind2="">
<subfield code="a">10 p. ;</subfield>
<subfield code="c">20 cm .    </subfield>
</datafield>
[...]
<datafield tag="651" ind1="" ind2="0">
<subfield code="a">United States</subfield>
<subfield code="x">History</subfield>
<subfield code="y">French and Indian War , 1755-1763</subfield>
<subfield code="v">Poetry .              </subfield>
</datafield>
<datafield tag="651" ind1="" ind2="0">
<subfield code="a">Louisbourg ( N.S.)</subfield>
<subfield code="x">History</subfield>
<subfield code="y">Siege , 1758</subfield>
<subfield code="v">Poetry .            </subfield>
</datafield>
<datafield tag="655" ind1="" ind2="7">
<subfield code="a">Poems</subfield>
<subfield code="y">1775.</subfield>
<subfield code="2">rbgenr      </subfield>
</datafield>
<datafield tag="752" ind1="" ind2="">
<subfield code="a">United States</subfield>
<subfield code="b">Rhode Island</subfield>
<subfield code="d">Newport .         </subfield>
</datafield>
<datafield tag="830" ind1="" ind2="0">
<subfield code="a">Early American imprints .</subfield>
<subfield code="n">First series ;</subfield>
<subfield code="v">no . 14254.           </subfield>
[...]
</record>
```

Each bibliographic record starts with the *administrative* metadata, which consists of a leader field, followed by several control and classification fields (data fields starting with 0XX). In the control and leader fields, several pieces of infor-

[26] ISO stands for International Standards Office; ISO 2709 is a standardized format for bibliographic descriptions.

[27] This example listing was shortened, for display purposes, for inclusion in this thesis. It does, however, reflect the main data fields of a MARC21 record. Most of the fields excluded from this example contained duplicate and/or redundant metadata. The MARC21 record structure is elaborate, and allows for inclusion of the same metadata in a variety of fields (Zhang, 2006).

mation are encoded into a fixed-length string, which can be used to check and uniquely identify an imprint. For example, in control field #008, the first six positions denote the creation date of the record: May 20th, 1985.[28] The trailing 's1775' signifies that a single publication date is known for this imprint: the year 1775. Also included in this fixed-length string are language information (eng for English), as well as other production information. The main data fields, however, start from #100 onward. [29]

The administrative metadata is followed by the *bibliographic metadata*, which starts from data field #100. Data field #100 holds information about a person connected with the work described in the record, in this case the author. The full name of the author is found in subfield #100-a, along with the dates associated in this name in #100-d, in this case the year of birth and suspected year of death.[30] The next important piece of metadata can be found in data field #245, which holds the title (a), subtitle (b), a 'statement of responsibility' (c), and the medium/resource type (h). Fields #250 through #270 hold edition, imprint, and other publication information, whereas data fields starting with 3XX contain physical information about the imprint: along with the imprint's physical dimensions these can also be playing time, trade price, and geospatial reference data (geographical coordinates), if available and applicable.

The Subject Access Fields occupy the range starting with 6xx. The data field entries with #651 denote a 'subject added entry' with a geographic name as the (main) entry element: Geographic Names (subfield a: United States, Louisburg), Form Subdivisions (v: Poetry), General Subdivisions (x: History), and Chronological Subdivisions (y: French and Indian War, Louisburg Siege). In data field #655 we can find terms that indicate Genre and/or Form, in this case 'Poems' ('rbgenr' references the specific classification scheme used: the ACRL Genre Terms thesaurus for rare books and special collections; 1775 again refers to the publication year).

The imprint's publication location can be found in data field #752, in a hierarchical format: Country (a), First-order Political Jurisdiction (b: state or province), Intermediate Political Jurisdiction (c: county or municipality), and City (d). Lastly, series or collection information is described in field #830: this imprint is part of the Early American Imprint Series, under 'sequential designation' (unique identifier) 14254.

In summary, the MARC21 records contain various types of bibliographic metadata on the imprints, varying from metadata describing the imprint's properties (unique identifier, title, subtitle, year, physical format) to metadata describing

[28] This data about a bibliographic record could be construed as meta-metadata.

[29] A full list of field and subfield codes can be found at `http://www.loc.gov/marc/bibliographic/`.

[30] Accordingly, the metadata on other entities as creator of the imprint can be found in data fields #110 (Corporate entities) and #111 ('Meeting' entities, also known as Conferences (See section 3.2.5)).

entities associated with the imprint, along with its relation to the imprint: People (as authors, or publishers, perhaps even as subjects), organizations or companies (again, as authors or publishers), Topic (as subjects, with general and specific subdivisions), Locations (as subjects (geographical coverage), publication location), Genres/Forms, Language.[31] The records are organized per imprint: each imprint has one record, and this means that some authors, subjects and locations are repeated oftentimes, thus creating redundancy and the potential for error (wrong input, misspellings, variations).

In recent years, a discussion has commenced on whether modern-day library catalog systems should move away from a relatively old—some say outdated—document-centered single entity data format, to a model/format which makes better use of current technology and conventions. This can be done by making use of a model which describes a publication as the function of a work, its expression into an imprint, and its relation to other entities (various agents as its creator, publisher, as well as other entities as places/locations, topics as subjects, etc.) (MARC, 2006). One of the major initiatives in this direction is the Functional Requirements for Bibliographic Records (FRBR) initiative proposed by the International Federation of Library Associations and Institutions (IFLA) (IFLA, 1998).[32]

3.2.2 Entity-relation-oriented: Data Tables and Semantic Triples

The FRBR data model proposed by the International Federation of Library Associations and Institutions exemplifies the move from a document-centered single-entity 'flat' data format toward a multi-entity hierarchical or networked model. If we want to implement this model into an on-line finding aid, there are two major data storage/description models that are commonly used: the *relational* (database) model (introduced in the 1970s by Edgar F. Codd) and the *object-entity* model. The SWHi project makes use of the latter, and for that reason we will go into such a data model in further detail. The relational model will be used as reference, however.

In the following section we will explain the transformation from the document-centered MARC21 metadata format into the *Semantic Triples* used by the Se-

[31] The relationships of the entities to the imprint (author, publisher, genre, subject, language) correspond largely with the taxonomical classification of the *Evans Digital Edition*, which was also created by using this MARC21 metadata source.

[32] FRBR further distinguishes between works, and manifestations and expressions of works (IFLA, 1998). In the case of *Evans*, the different imprints which denotes variations or translations of the same work, would—in FRBR—be regarded as manifestations of this work. As this information was not available in the original MARC21 metadata, and the emphasis of the entire *Evans* initiative has been on early American *imprints* instead of works, we have chosen not to try to follow the FRBR guidelines to this detailed level.

mantic Web for History project. We will illustrate this normalization and transformation process from the perspective of the Information Visualization Reference Model proposed by Card et al. (1999).[33]

Data Tables

Table 3.1 shows the initial data table for imprint no. 14254, directly based on the MARC21 metadata shown in listing 3.1. This data table is not yet normalized: it contains multiple instances of the Subject metadata field. Such a data table can also contain multiple instances of the Author Field, or multiple instances of one person in separate roles, for instance if John Maylem would be both Author and Publisher. Furthermore, if John Maylem also appears in other imprint records, this would also cause redundancy. In order to eliminate this redundancy, we need to further transform the data tables until we have a (normalized) data table for each entity type (Person, Location, Topic, etc.).

In the next step of the transformation cycle, all separate entities are extracted from the main imprint data table, and each entity type is placed in its own derived data table. Tables 3.2 and 3.3 show separate data tables for this next normalization step. In table 3.2, each of the two persons, who previously appeared in the imprint record as mere metadata properties of the imprint, is promoted to a separate entity instance, each with its own data table record. Please note that this data table can be transformed and normalized even further: in a large imprint collection like Evans, people can appear in the metadata in various roles: as author, co-creator, translator, printer, et cetera. Also, a person can have one or more interpersonal relation (co-creator of an imprint with, ancestor of, friend/acquaintance of) to other people. That same person can also have connections to multiple imprints, or links to several events or locations. These one-to-many and many-to-many relations between entities require further transformations and normalization steps.[34]

[33] The Information Visualization Reference Model offers a framework in which '*raw data*' (data in an idiosyncratic format), by a (cyclic) series of normalization and transformation steps, is transformed into a '*data table*'. These data tables are relational descriptions of data extended to include metadata (and can also contain (relational/semantic/hierarchic) network structures). Each row in the data table denotes a metadata field. Each field consists of an identifier (variable name), a data type (Nominal, Ordinal, Interval, Quantitative), and a value. The Quantitative data types are subdivided further: into Quantitative 'proper' (Q), Quantitative Temporal (Qt), and Quantitative Geographical (Qg). The data tables based on these data types, especially the various quantitative dimensions, form the basis for further transformation into 'visual structures' and 'views', but this lies beyond the scope of this section (Card et al., 1999; Card, 2003). This reference model also be used as reference in the next chapter, when we will be discussing the role of Information Visualizations in effective interfaces. See chapter four.

[34] In the relational model, describing this relation would require the use of additional tables and foreign keys. If we were to draw a diagram the entities and its relations would result in a network diagram. See also figure 3.7 and section 4.3.3 (Network Graph Visualizations)

EvansID	*N∗*	14254
Title	N	The conquest of Louisburg
Author	N	John Maylem
OriginalYear	Qt	1758
Genre	N	Poems
Publisher	N	Solomon Southwick
PubPlace	N	Newport, RI
PubYear	Qt	1775
NumPages	Q	16
Format	N	16mo
Subject	N	United States, History ..
Subject	N	Louisburg (N.S.), History ..
Description	N	..
..	..	..

Table 3.1: Initial Data Table for Evans Imprint Metadata. *N=Nominal, Q=Quantitative, Qt=Quantitative Temporal.

PersonID	*N∗*	*MaylemJohn*	*SouthwickSolomon*
FullName	N	John Maylem	Solomon Southwick
FirstName	N	John	Solomon
LastName	N	Maylem	Southwick
EntityType	N	Person	Person
YearBirth	Qt	1739	-
YearDeath	Qt	1762?	-
Role	N	Author	Printer
..	..	..	..

Table 3.2: Person Entity data table derived from the initial Imprint data table shown in Table 3.1. *N=Nominal, Qt=Quantitative Temporal.

Table 3.3 shows a data table of the locations derived from our example imprint. Both city locations have a relation to the same imprint, albeit a different one. The location metadata taken from the MARC21 metadata has contained within it a hierarchical structure of consisting of three levels: city, state/province, and country.[35] Besides the cities we can also derive four other location entities from the original metadata record (Rhode Island, Nova Scotia, the United States, Canada). This requires a further transformation of the location data entities of

[35] In a historical context, using names of states and countries, and current other geopolitical subdivisions to denote historical locations are of course anachronistic. The states/provinces and countries used here are taken directly from the MARC21 metadata records, which uses modern (current) geopolitical entities to describe locations. In the reality of the 1758 siege, Louisb(o)urg was part of French North America before, and of British North America after the siege. See also section 4.3.3 (Maps).

LocID	*N*∗	*NewportRI*	*LouisburgNS*	..
Name	N	Newport	Louisburg	..
EntityType	N	Location	Location	..
LocType	N	City	City	..
State	N	Rhode Island	Nova Scotia	..
Country	N	United States	Canada	..
..	..	..	..	

Table 3.3: Location Entity data table derived from the initial Imprint data table shown in Table 3.1. *N=Nominal, Qt=Quantitative Temporal.

table 3.3 into separate advanced data tables for each of the hierarchic subdivisions of the Location entity type. One such advanced data table (table 3.4) shows the state of Rhode Island as a further hierarchical subdivision. Data tables can also be augmented with additional data, such as data derived from data aggregations (Total Imprints, derived from the full *Evans* metadata collection), or information derived from other sources (Year of First Press, derived from the back pages of Charles Evans' *American Bibliography*, and geospatial coordinates, derived from the Google Map API).

State	*N*∗	*RI*	..
Name	N	Rhode Island	..
EntityType	N	Location	..
LocType	N	State	..
YrFirstPress	Qt	1727	..
TotalImprints	Q	1678	..
Latitude	Qg	41.7	..
Longitude	Qg	–71.5	..

Table 3.4: Location data table, at state level: an advanced data table for Evans Metadata. *N=Nominal, Q=Quantitative, Qt=Quantitative Temporal, Qg=Quantitative Geographical.

Semantic Triples

The Semantic Web is a vision of a future Web, envisioned by Tim Berners Lee, the inventor of the World Wide Web. One of the its prominent features is the potential for interoperability and reuse between ontologies, which allows the automated processing and interoperation between agents referenced earlier (in the introduction) (Berners-Lee et al., 2001; W3C, 2001). A data collection, such as *Evans*, is stored and defined in an *ontology*.

All ontologies are defined within strictly described namespaces, and a particular ontology can make use of (combinations of) commonly used—commonly agreed on—namespace descriptions or ones defined in earlier efforts, or can make extensions to this by defining its own *namespace*. Common semantic web namespaces are that of XML, RDF, RDF-schema, proton, DC (an implementation of the Dublin Core metadata description framework), FOAF (social networks) Vicodi and HEML (both specific for historical purposes).

In RDF, information is a collection of statements, each with a subject, predicate (sometimes called verb), and object. This simple—yet expressive—three-part data structure is called a triple. All entities, properties, and relations can be stored in this way (*subject–predicate–object*). For example, an imprint (subject) has as its creator (predicate) a specific person (object). Restated: Imprint no. 14254—'hasCreator'—John Maylem.[36]

By explicitly defining *and* formally describing the relations between objects in the ontology, these relations get an additional semantic meaning. In other commonly used data storage models, like the Relational Database Model, each entity is treated differently, and stored separately.[37] In a relation database, comprised of *tables* and links—or *relations*—each entity is stored in a separate table. Each table has its own structure and its own properties ('fields"'); the table 'Persons' for instance can have four fields(first name, last name, year of birth and year of death), whereas the table 'Imprints' may have ten or more. Furthermore, the relations linking all tables are not defined explicitly. All that is defined is which fields of which tables are linked to one and another, but what the (semantic) meaning of this link/relation denotes remains explicit. Also, when adding new relations to a relational database, the data structure of several tables needs to be altered, additional (translation) tables need to be created and populated, whereas in ontology the new relation only needs to be defined in the namespace (as a triple), and be put in place, also as a triple, relating the two entities to each other. Such a semantic structure can best be demonstrated by an example (see listing 3.2).

Listings 3.2 and 3.3 show the *Evans* metadata which has been transformed from MARC21 into a set of semantic RDF triples (in this only one imprint—*The Conquest of Louisburg*—and one person—John Maylem—are shown; the other entities and their triples are expressed in the same format). Every line in this example denotes a triple: the first triple, which has 'Imprint-14254' as a subject, consists of the triple 'Imprint-14254'—'dc:language'—'#langEnglish': Imprint no. 14254 is in the English language. English, as a language, is formulated here as a separate object (denoted by 'rdf:resource'), and has its own set of triples.

The second set of triples shown in the example are for the author of imprint no. 145254 (listing 3.3). According to the SWHi Ontology describing the *Evans* collection, there are sixteen triples which have John Maylem as a subject. However,

[36] Restated in a less cryptic way, the predicate describes 'hasCreator' the relation 'has as its creator' or 'is created by'.

[37] We choose to offset the Semantic Web triple structure against that of the Relational Database Model because it is the most commonly used model to store intricately structured data, both on and off the Web.

there are also more entities that have a connection to the person John Maylem, that are not shown in this listing. In the *Evans* collection, Maylem is shown to be the author of three imprints in total. In the ontology, this is evidenced by three instances where #prsMaylemJohn is an *object*.

Listing 3.2: Ontology instances for Imprint no. 14254 (RDF-OWL XML format)

```
<swhi:Poems rdf:ID="Imprint-14254">
  <dc:language rdf:resource="#langEnglish" />
  <dc:publisher rdf:resource="#prsSouthwickSolomon1775" />
  <dc:coverage rdf:resource="#locUnitedStates" />
  <rdfs:label>The conquest of Louisburg</rdfs:label>
  <dc:description><![CDATA[Imprint supplied by L.C. Wroth in &quot;John Maylem: poet and
   warrior .&quot; Publications of the Colonial Society of Massachusetts, v. 32 (1937):
   p. [87]-120.]]></dc:description>
  <swhi:originalPublicationInformation>[Newport, R.I. : Printed by Solomon Southwick, 1775]
  </swhi:originalPublicationInformation>
  <swhi:publishedIn rdf:resource="#locUnitedStatesRhodeIslandNewport" />
  <dc:source rdf:resource="http://opac.newsbank.com/select/evans/14254" />
  <dc:subject rdf:resource="#sbjFrenchAndIndianWar1755-1763" />
  <dc:description>Caption title.</dc:description>
  <dc:relation>Evans 14254</dc:relation>
  <dc:creator rdf:resource="#prsMaylemJohn1739-1762" />
  <dc:subject rdf:resource="#sbjLouisbourgNSHistorySiege1758Poetry" />
  <swhi:authorTitle>By John Maylem, philo-bellum</swhi:authorTitle>
  <dc:subject rdf:resource="#sbjUnitedStatesHistoryFrenchAndIndianWar1755-1763Poetry" />
  <dc:subject rdf:resource="#sbjSiege1758" />
  <rdf:type rdf:resource="&protont;Document" />
  <dc:format>Electronic text and image data. [Chester, Vt. : Readex, a division of
   Newsbank, Inc., 2002-2004. Includes files in TIFF, GIF and PDF formats with inclusion
   of keyword searchable text. (Early American imprints. First series ; no.
   42884).</dc:format>
  <dc:coverage rdf:resource="#locLouisbourgNS" />
  <swhi:subTitle>A poem. /</swhi:subTitle>
  <dc:subject rdf:resource="#sbjUnitedStates" />
  <dc:format>Microform version available in the Readex Early American
   Imprints series.</dc:format>
  <dc:title>The conquest of Louisburg</dc:title>
  <dc:relation>Alden, J.E. Rhode Island, 609</dc:relation>
  <dc:subject rdf:resource="#sbjHistory" />
  <dc:relation><![CDATA[Shipton &amp; Mooney 42884]]></dc:relation>
  <dc:format>Electronic text and image data. [Chester, Vt. : Readex, a division
   of Newsbank, Inc., 2002-2004. Includes files in TIFF, GIF and PDF formats with
   inclusion of keyword searchable text. (Early American imprints. First series ;
   no. 14254).</dc:format>
  <dc:format>10 p. ; 20 cm.</dc:format>
  <dc:description>Evans entry 14254 records a 16 p. edition with imprint: Boston, Printed
   in 1758. Newport (R.I.) Reprinted by Solomon Southwick, in 1775. According to Alden,
   that edition is a nineteenth century reprint.</dc:description>
  <dc:subject rdf:resource="#sbjLouisbourgNS" />
  <dc:relation>Bristol B4058</dc:relation>
  <dc:subject rdf:resource="#sbjPoetry" />
  <dc:date rdf:resource="#t1775" />
</swhi:Poems>
```

Listing 3.3: Ontology instances for person John Maylem (RDF-OWL XML format)

```
<protont:Person rdf:ID="prsMaylemJohn1739-1762">
  <foaf:interest rdf:resource="#sbjLouisbourgNS" />
  <foaf:interest rdf:resource="#sbjSiege1758" />
  <swhi:mentionedIn rdf:resource="http://opac.newsbank.com/select/evans/8193" />
  <swhi:mentionedIn rdf:resource="http://opac.newsbank.com/select/evans/14254" />
  <foaf:interest rdf:resource="#sbjUnitedStates" />
  <foaf:interest rdf:resource="#sbjUnitedStatesHistoryFrenchAndIndianWar1755-1763Poetry" />
  <foaf:interest rdf:resource="#sbjHistory" />
  <swhi:mentionedIn rdf:resource="http://opac.newsbank.com/select/evans/8194" />
  <foaf:family_name>Maylem</foaf:family_name>
  <rdfs:label>Maylem, John, 1739-1762</rdfs:label>
  <foaf:interest rdf:resource="#sbjPoetry" />
  <foaf:interest rdf:resource="#sbjLouisbourgNSHistorySiege1758Poetry" />
  <foaf:firstName>John</foaf:firstName>
  <swhi:exists rdf:resource="1739-1762?" />
  <foaf:interest rdf:resource="#sbjFrenchAndIndianWar1755-1763" />
  <foaf:name>John Maylem</foaf:name>
</protont:Person>
```

The semantic network or 'web' resulting from this collection of triples is (partly) reconstructed in figure 3.7. Imprint no. 14254, *The Conquest of Louisburg*, is linked (via the arrows, representing the predicates) to other entities. John Maylem, the author of Imprint-14254, is also the author of Imprint-8193.

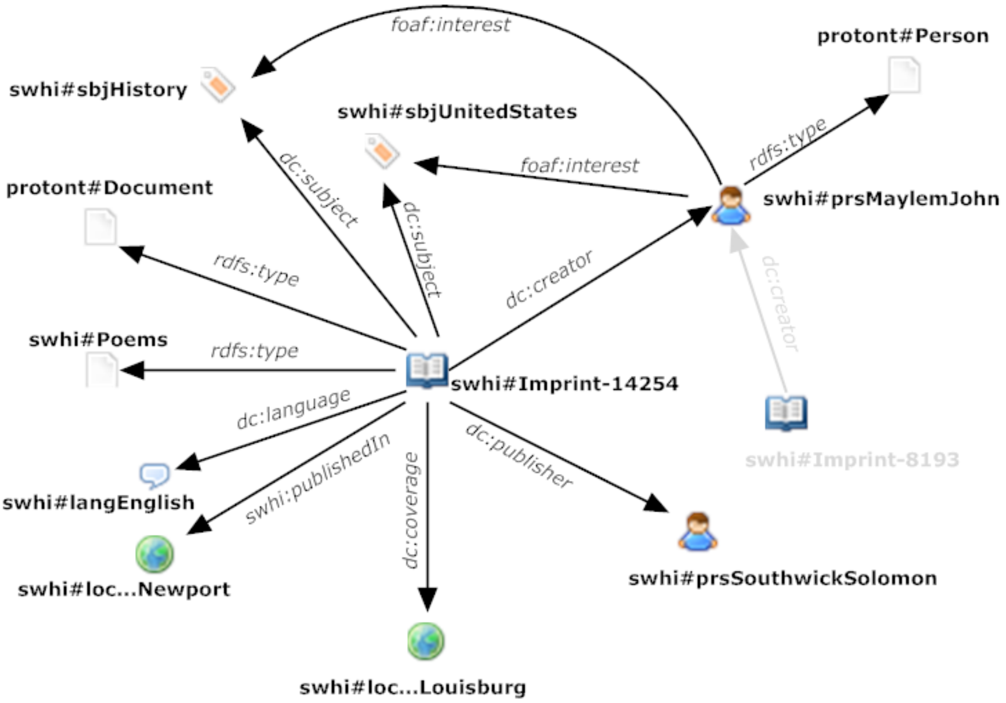


Figure 3.7: A Network visualization of (part of) the SWHi semantic triples, taken from listings 3.2 and 3.3. Entity instances (persons, imprints, locations, etc.) can be subjects and objects in a semantic triple: the direction of the arrow (which represents the predicate) goes from subject to object.

Instead of a one-dimensional—'flat'—metadata structure describing only documents(like in MARC21), the *Evans* metadata is transformed to a multi-dimensional entity-relation-based information space (see figure 3.7). The 35,823 imprints of the *Evans* collection have been transformed into over 80,000 separate entities (see table 3.5), with over two million triples describing and interlinking these entities.

The resulting Semantic structure is not only of great use for software agents and for automatic processing, but also for 'human agents': people (e.g. historians) who are interested in this new Semantic network of people, the events they are involved in, and of the documents they have written. This semantic network is about meaningful relations, and the more meaning that can be ascribed to these relations, extracted from the source data, the more meaningful this semantic framework becomes for researchers.

How exactly these relations—and more importantly the additional relations we can infer through them—can be of use to (historical) researchers can be witnessed in chapter five, where Semantic inferencing is applied to create a powerful semantical similarity mechanism[38] to aid and support scholars with their information-seeking in an on-line information space.

ResourceType	*numberofentities*
Document	35,823
Person	23,688
Subject	12,828
TimeInterval	4,042
Organization	2,582
Location	1,787
Event	296
Language	21

Table 3.5: Contents of the SWHi ontology: Main resource types and their frequencies.

3.3 Chapter Evaluation

In this chapter we have followed the imprints of the *Evans* collection and its metadata from the early twentieth century into the twenty-first, from bibliographies and index cards into the digital age. Charles Evans has provided the world of historical scholarship with a lasting and impressive legacy. The analysis of Evans' *American Bibliography* and its (analog and digital) successors has uncovered a valuable taxonomy of the collection's structure on which we can base the next on-line incarnation of *Evans*: the most important entry point into the Evans collection are Authors, Printers/Publishers (an example of the provenance-based name-collecting behavior of historians), Locations, Genres, Subjects, Eras (temporal subdivisions) and Languages. Charles Evans (and his successors) note the special importance archivists and historians ascribe to *dates*, *eras*, and *chronology*, as can be evidenced both explicitly (in Evans' preface) and implicitly (by the chronological arrangement of most of the bibliographies and finding aids).

The taxonomical and historical analysis has also uncovered another important fact: that a wide variety of historians was—and still is—interested in this resource. This wide variety of audiences (different groups of historians, archivists, etc.) did show the downsides of the print format, though: each audience had its own preferences for accessing the collection, and each audience had its own part of the collection that they deemed most important.

The organization of a bibliography in book format is not flexible, however. As a user, one is bound by the organizational decisions made by the bibliographer. The user cannot arrange the information in the way that is most convenient to his purposes. Furthermore, finding a particular imprint is hard, when one does not

[38] See section 5.1.1.

know the specific year. Finding imprints on a specific subject, without knowing year, title or author can be a tedious and time-consuming undertaking with a bibliography in book format. Charles Evans did, however, realize this and partly compensated for this by adding separate brief indices for Author, Genre, and Location, at the back of each volume of its American Bibliography.

The advent of new technology came with a solution to these problems, and as the *Evans* collection's metadata was transformed into a digital format, and eventually an electronic version of *Evans* became available on the web. The resulting on-line finding aid, the *Evans Digital Edition* allows for much more flexibility than had ever been possible with the analog bibliographies. In an on-line environment, the user can now rapidly access specific subsets of interest, change the order of listings to what is most convenient to him. With an on-line electronic finding aid the historian is not restricted to one single primary access point anymore, and also not bound to one predetermined ordering mechanism.

Further developments in data modeling have also brought more flexibility the way metadata is defined and stored. Whereas the MARC21 bibliographic format only focuses on the imprint, functional, object-oriented models and initiatives like FRBR and the Semantic Web move away from document-centeredness, in favor of a multi-faceted approach. This multi-faceted approach focuses on various entities associated with bibliographic data, like imprints, people, locations.

The transformation of metadata into semantic triples is a rather novel way of modeling bibliographic data. It not only stores all properties of the entities, it also explicitly defines the relations between these entities. In semantic triples all metadata can be stored in a simple, yet powerful fashion. It provides a meaningful and rich definition of an imprint, its properties, and all connected entities.

Chapter 4

Historical Interface Design

> 'The working mind is greatly leveraged by interaction with the world outside it' (Card, 2003)

Historians search and find information differently than other users. They display information-seeking behavior that differs from the traditional IR model. Still, the vast majority of finding aids available for scholars—library catalogs, search engines, digital collection repositories—seem to treat all users in the same way by focusing solely on traditional Information *Retrieval*. The focus of most tools available for scholarly research lies not on user requirements nor on the specifics of the data source: the focus lies on custom and on technology.

In this chapter we attempt to provide an alternative to this one-size-fits-all approach. With the information gathered from the previous chapters we will try to answer the following research questions:

- What constitutes an ideal electronic finding aid for scholarly use by historians?
- Which interface elements and system features can be used to build such a finding aid?
- And what role can Information Visualization play in such an interface?

These research questions can be summarized as follows: what should a search interface for use by historians look like? In this chapter we attempt to couple the user to our data, via an interface.

We will first summarize the requirements derived from our findings of the previous chapters. After this, a conceptual model for a historical berrypicking interface is introduced, Cole's 'name-drawer schema'. From this model we derive the elements that we think should go into an on-line electronic finding aid.

We will look at the architecture and the interface elements we think should constitute such a finding aid. The data and its taxonomy can serve as the framework and the building blocks for the interface elements, whereas our knowledge about the user will help us make important design decisions. A special focus will be directed toward the role of Information Visualization in such a 'berrypicking' interface, centered around sensemaking, context-building and discovery.

4.1 Coupling User to Data: Requirements

What constitutes and ideal electronic finding aid? The first of the research questions can be answered—at least partially for the moment—from the previous chapters. From an analysis of the user (the historian: chapter three) and the data (*Evans*: chapter three) we can derive a set of requirements for an electronic finding aid for historical primary source data. The finding aid should:

- be built to specifically accommodate for the historian's Information-seeking Behavior, centered around name-collecting and berrypicking;
- enable and empower the cognitive aspects of this IB: item/pattern discovery, context-building, impression formation;
- support both formal (searching) and informal (browsing, serendipity) Information-seeking methods;
- be flexible to accommodate for individual preferences and the IB of multiple audiences[1]: offer multiple access points, multiple sorting mechanisms, multiple ways to view the data;
- provide access to the source (as proof and evidence).

The focus of such a system should be on ease of use: it should be easy to use and understand. Historians are not necessarily computer experts, and they might value computers and digital tools differently than the average library or computer scientist (as developers). Furthermore, historians are primarily source- and text-oriented, perhaps not as apt and familiar with advanced visualizations as other (e.g. science) scholars. This should be kept in mind when evaluating and implementing advanced features: Advanced features may be implemented, and insight into the underlying (semantic) data structure may be given, but historians are not likely to be interested in these features. Therefore, the focus should not be on what is possible technologically, but on what might be beneficial to the user. The end users are historians, not computer scientists.

[1] E.g. print historians, economic historians, who each have their own point of interest within the data set, and who might also have different preferences for viewing this data.

4.2 Interface Elements

> ‘Humans have remarkable perceptual abilities that are greatly underutilized in current designs’ (Shneiderman, 1996)

As we have noted earlier, traditional electronic finding aids are focused almost solely on traditional Information Retrieval (keyword searching). Furthermore, they are document-centered and do not offer the flexibility users might expect, especially when other on-line (non-scholarly) resources seem to become more and more user-centered.

In such an environment many users—including historians—are hindered when they attempt to perform their basic information-seeking tasks (orientation, seeking known material, context-building, and identifying relevant material (Duff and Johnson, 2002)).[2] We propose a number of additions to electronic finding aids, in the form of various visual aids, both textual and graphical: all of these interface elements exist in some form on the World Wide Web, but they seem to be missing from nearly all electronic finding aids for scholarly information.

4.2.1 Browsing

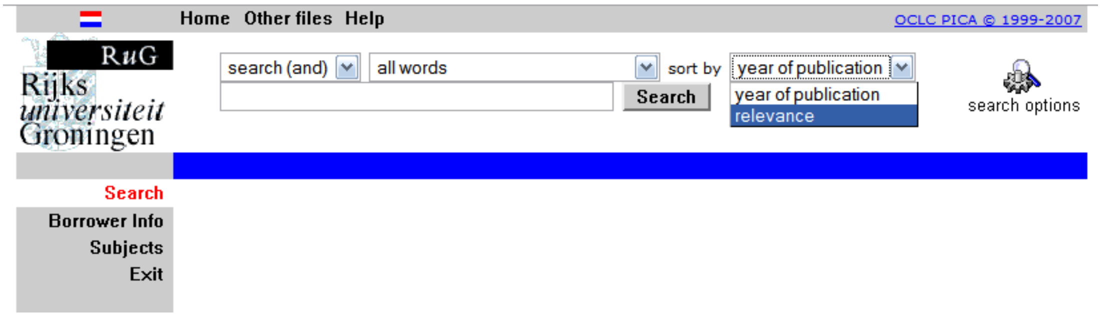
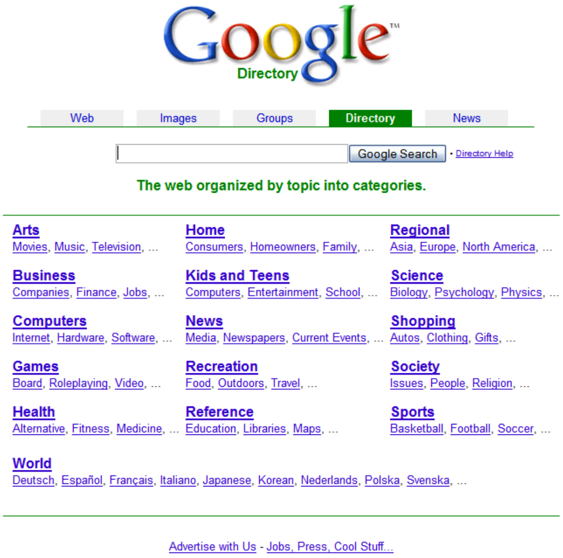

search (and) all words sort by year of publication

Figure 4.1: Library catalogs cater mostly to Information Retrieval via *searching*. The starting page of Library Catalog of the University of Groningen. Source: `http://opc.ub.rug.nl`.

There are two primary ways of navigating information spaces on the web: *searching* and *browsing*. Even though most humanities scholars and social scientists—including historians—seem to prefer informal methods of information-seeking, many scholarly finding aids offer little opportunity for browsing and serendipitous discovery of items. Although searching and browsing are not mutually exclusive, most on-line library catalogs’ starting pages seem to exist of little more than a search input screen (see figure 4.1), and thus almost ‘force’ the user to search instead of browse. Pandit and Olston (2007) cite three reasons why people prefer browsing over searching as a information-seeking strategy:

[2] See also section 2.2.2 (The historian’s information-seeking behavior).



The web organized by topic into categories.

Arts
Movies, Music, Television, ...

Business
Companies, Finance, Jobs, ...

Computers
Internet, Hardware, Software, ...

Games
Board, Roleplaying, Video, ...

Health
Alternative, Fitness, Medicine, ...

Home
Consumers, Homeowners, Family, ...

Kids and Teens
Computers, Entertainment, School, ...

News
Media, Newspapers, Current Events, ...

Recreation
Food, Outdoors, Travel, ...

Reference
Education, Libraries, Maps, ...

Regional
Asia, Europe, North America, ...

Science
Biology, Psychology, Physics, ...

Shopping
Autos, Clothing, Gifts, ...

Society
Issues, People, Religion, ...

Sports
Basketball, Football, Soccer, ...

World
Deutsch, Español, Français, Italiano, Japanese, Korean, Nederlands, Polska, Svenska, ...

Advertise with Us - Jobs, Press, Cool Stuff...

Figure 4.2: The Google Directory, powered by the Open Directory project, offers the user a way to 'orienteer' and get a sense of overview. Web pages on a specific subject can be found by browsing a hierarchical category structure. Source: `http://www.google.com/Top/`.

1. difficulty in formulating appropriate queries (which keywords to use for traditional IR query)

2. open-ended search tasks (which often entail a significant amount of manual navigation, as part of an extended process of exploration, discovery, and task/query refinement: see also HIB and berrypicking)

3. preference for 'orienteering' (rather than teleporting, this enables them to understand the surrounding context)

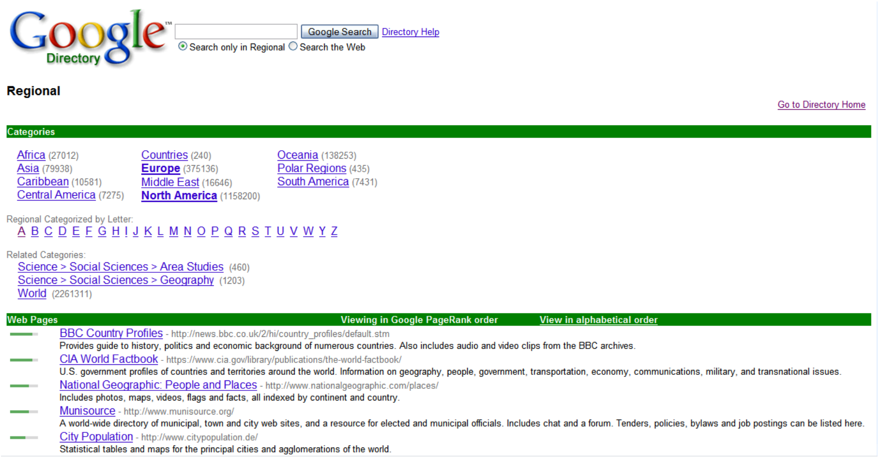
Unfortunately for many, this search-only approach, which seems prevalent on the web and with library catalogs, is often the only way to access bibliographical data. When employing a keyword search users only see two levels of a collection: the top node (the entry page with the search box) and bottom node (the results page and the imprint detail page); all sense of overview is lost.

Just like with on-line shops—and with traditional archives[3]—people prefer to 'orienteer', look around first, see what else is there, take a sample from the shelves, to get a sense of overview, to see 'what's in store'. (Pandit and Olston, 2007; Duff and Johnson, 2002; Delgadillo and Lynch, 1999; Rivadeneira et al., 2007)

If such an overview is missing, it is hard to get a sense of which types of publications are in a collection, what the highlights are, and perhaps even most

[3] Historians prefer to get a sense of overview first, 'rummage around', browse the shelves, ask the archivist for interesting items in the collection. See section 2.2.2

importantly where to start. Offering the users a way to hierarchically browse the collection is perhaps the most obvious solution here: users can locate their desired resource more easily if they are organized hierarchically. This is especially true for large datasets (Li et al., 2007). The *'catalog' or directory model* provides a way to browse a (bibliographic) collection by presenting the data in a hierarchical format. This model was popularized on the Web by the Yahoo! Directory, which originally started out in 1994 as an hierarchically organized access point to the web.[4] Other well-known applications of the directory model are the Google Directory[5] (see figures 4.2 and 4.3) and the Open Directory project.[6] The current version of *Evans Digital Edition* also makes use of this hierarchical browsing model, as can be witnessed in figure 3.5, in the previous chapter.



Regional

Go to Directory Home

Categories

Africa (27012)	Countries (240)	Oceania (138253)
Asia (79938)	Europe (375136)	Polar Regions (435)
Caribbean (10581)	Middle East (16646)	South America (7431)
Central America (7275)	North America (1158200)	

Regional Categorized by Letter:
A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Related Categories:
Science > Social Sciences > Area Studies (460)
Science > Social Sciences > Geography (1203)
World (2261311)

Web Pages Viewing in Google PageRank order View in alphabetical order

BBC Country Profiles - http://news.bbc.co.uk/2/hi/country_profiles/default.stm
Provides guide to history, politics and economic background of numerous countries. Also includes audio and video clips from the BBC archives.

CIA World Factbook - https://www.cia.gov/library/publications/the-world-factbook/
U.S. government profiles of countries and territories around the world. Information on geography, people, government, transportation, economy, communications, military, and transnational issues.

National Geographic: People and Places - http://www.nationalgeographic.com/places/
Includes photos, maps, videos, flags and facts, all indexed by continent and country.

Munisource - http://www.munisource.org/
A world-wide directory of municipal, town and city web sites, and a resource for elected and municipal officials. Includes chat and a forum. Tenders, policies, bylaws and job postings can be listed here.

City Population - http://www.citypopulation.de/
Statistical tables and maps for the principal cities and agglomerations of the world.

Figure 4.3: The Google Directory. This on-line hierarchical catalog allows for the items (in this case web pages) to be listed at any node of the hierarchy, as well as the items to be ordered and subdivided in multiple ways. Source: `http://www.google.com/Top/`.

A hierarchical browsing model can easily be combined with a search mechanism, to allow the user to search either within the selected node (directory level) or within the entire collection.

We propose a catalog which is on a combination of both the hierarchical catalog structure of the *Evans Digital Edition*[7], and the semantic object entities identified in the previous chapter.[8] Such a combined approach—in our view—goes together well with the provenance-based orientation of the *Digital Edition* and the historian's name-collecting, while emphasizing the dual role of the tax-

[4] Yahoo! was originally an acronym that stood for 'Yet Another Hierarchical Officious Oracle'. It did not become a regular search engine until much later. (Source: `http://docs.yahoo.com/info/misc/history.html`). The current Yahoo! Directory can be accessed at `http://dir.yahoo.com/`.

[5] `http://www.google.com/Top/`.

[6] `http://dmoz.org/`.

[7] See section 3.1.5.

[8] See section 3.2.2.

onomy's principal entities (which act as pivotal points for both architecture and Information Visualizations).[9] This dual role applies to *people* and *organizations* (which can be either authors, creators, publishers, or the topical subject of an imprint) as well as to *location* (which denote the place of publication as well as the topical geographic coverage of an imprint). The categories and pivotal entities for the catalog proposed are:

1. People and Organizations (Provenance-based: subdivided into listings for these 'agents' in their various roles)
2. Locations (Geographical: subdivided primarily into three geographical levels: Country, State, City, with a concurrent option to browse 'by role')
3. Genres (or Document Types)
4. Subjects (Topical: with crosslinks to People/Organizations and Locations)
5. Events and Eras (Chronological: also includes a hierarchical subdivision into centuries and decades)
6. Languages

A hierarchical browsing structure as proposed above—in our opinion—fits very well exploratory open-ended searching favored by historians. Orienteering, serendipity and item discovery can be further supported and enhanced by making use of frequency-ranked lists[10] and employing shortcuts to the most important subcategory within each category (allowing the user to 'skip a node' (Shneiderman, 1997)).[11]

4.2.2 Searching

Even though browsing might seem a better match with the historian's information-seeking behavior, offering only a browsing mechanism will make finding information a possible herculean task for a researcher. For a large document set a keyword search interface is imperative. It enables closed-ended queries for specific term occurrences. Searching for known items[12] is also facilitated by offering both browsing and searching in an on-line finding aid environment.

Search interfaces come in two basic varieties: 'simple' and 'advanced'. In a simple search the user typically inputs the keyword terms into one free-text field,

[9] See chapter 3, introductory paragraphs.

[10] See also section 4.3.3.

[11] According to Figure 4.2 Google seems to favor listing three 'skip-a-level' shortcuts per category. Nielsen (2000) also favors a listings consisting of a small number of items to enhance scannability.

[12] See also section 2.2.2 (The historian's information-seeking behavior).

in a bag-of-words fashion. An advanced search allows the user to have much more control over which fields and content types to search, and over how the results are displayed. While advanced search interfaces provide more functionality, users tend to favor speed and simplicity to control: the simple search interfaces are preferred by most users (Fahmi et al., 2007).

A keyword search usually ends with a search engine result page (SERP), a listing of documents (or other entities)—along with other important information[13]—that match the keywords, ordered in a specific fashion. In a full-text Web search engine like Google[14], Microsoft Live Search[15] or Yahoo![16], the results are ordered by 'relevance': the most important results are listed first.[17] In a digital library environment, search results are ordered primarily by publication year, although the nature of the collection seems to dictate the exact order: library catalogs are typically ordered with the most recent publications listed first[18], whereas some historical document collections and archives tend to be ordered by default with the oldest record first.[19]. When we look at the arrangement of the *Evans* finding aids as well as at what is customary on the web, we should conclude that there are six possible sorting arrangements that can be implemented for the SERPs in our ideal finding aid:[20]

1. Chronological, (by year, by default: oldest first; alternatively: newest first)
2. Relevance (in the traditional IR sense: based on document vectors and the match between the keyword and the document content)[21]
3. Alphabetic (by title, A-Z)

[13] On a Web SERP page, the 'important' information is usually made up of the title of the result, a snippet of the page's content, a URL link, along with other metadata like page size and (index) date.

[14] `http://www.google.com`

[15] `http://www.live.com`

[16] `http://www.yahoo.com`

[17] Google uses an automated ranking mechanism they call 'Pagerank'. This ranking mechanism calculates the 'relevance' of a search result, by weighing the number of incoming links from other web pages, as well as the importance of the linking web pages. The exact formula of the Pagerank algorithm is a well-guarded business secret. Source: `http://www.google.com/technology/`.

[18] See for example the Dutch Royal Library catalog `http://opc4.kb.nl/`, the Harvard University library catalog `http://lms01.harvard.edu` and MIT's catalog `http://library.mit.edu`. Worldcat, the world's largest bibliographic database, however, lists the search results by 'relevance'

[19] See the *Evans Digital Edition* and the paper versions of the bibliographical finding aids to *Evans*. See also chapter three.

[20] Please note that the results returned by the search engine do not necessarily need to be restricted to documents only. All other object types we identified in chapter three can be returned as well, for instance as 'category matches'.

[21] See section 2.2.1: The traditional IR model.

4. By Provenance, Alphabetic (grouped by author, A-Z)
5. Relevance (similar to Google's pagerank: based on the relative importance of the object, i.e. the number of incoming links, connections, or (semantic, triple-based) relations.)
6. Popularity (based on the number of pageviews, or other user-determined metrics like rating or the number of times an object is tagged, saved or bookmarked)

The most likely candidates for determining a default sorting mechanism are sorting by chronology and by relevance. But in a complex multi-object type environment, what exactly constitutes 'relevance'? A mere keyword-text match may not yield the most important document first: listing the results most likely to be considered relevant by the user requires a more intricate and complex metric. We consider a 'hybrid' relevance metric, combining IR relevance, 'semantic relevance', and popularity (points 2, 5, and 6), to be the most promising solution.[22]

Further interesting approaches to user-centered SERP interfaces are parametric and faceted searches, which allow the user to filter, cluster and manipulate the search results (Shneiderman, 1997, 1998; Huynh et al., 2007; Tvarozek and Bielikova, 2007). The entities derived from the bibliographic metadata form an excellent basis to base the parameters and facets on. These faceted and parametric searches can be either textual or graphic.[23] Figure 4.4 shows a search engine results page with advanced sorting and filtering options, from the *Internet Archive*'s[24] text collection.

4.2.3 Berrypicking

One important—perhaps the most important—aspect of the historian's information-seeking behavior is that it bears similarity to Marcia Bates' Berrypicking model (Yakel, 2005; Duff and Johnson, 2002; Bates, 2005a; Bass and Rosenzweig, 2001). An interface that attempts to assist the historian in his quest for items that are relevant for his research, should therefore incorporate a berrypicking mechanism in some way or shape.

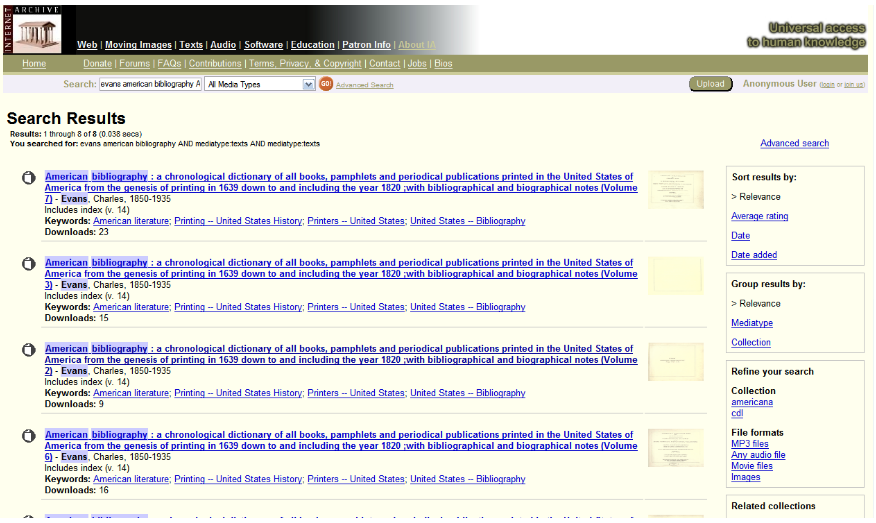
Cole's Name-Drawer Schema

In 2000, Charles Cole developed a interface model designed specifically to allow for berrypicking, mental contextualization, and pattern discovery. Cole called this model the 'names-drawer schema'(Cole, 2000a), a metaphor which he based

[22] The exact composition and relative weights of the three components will require further experimentation and user testing.

[23] See section 4.3 for a discussion of information visualizations.

[24] `http://www.archive.org`.


INTERNET ARCHIVE
Web | Moving Images | Texts | Audio | Software | Education | Patron Info | About IA
Universal access to human knowledge
Home
Donate | Forums | FAQs | Contributions | Terms, Privacy, & Copyright | Contact | Jobs | Bios
Search:
All Media Types
GO!
Advanced Search
Upload
Anonymous User (login or join us)
Advanced search

Search Results

Results: 1 through **8** of **8** (0.038 secs)
You searched for: evans american bibliography AND mediatype:texts AND mediatype:texts

Search: evans american bibliography A

American bibliography : a chronological dictionary of all books, pamphlets and periodical publications printed in the United States of America from the genesis of printing in 1639 down to and including the year 1820 ;with bibliographical and biographical notes (Volume 7) - **Evans**, Charles, 1850-1935
Includes index (v. 14)
Keywords: American literature; Printing -- United States History; Printers -- United States; United States -- Bibliography
Downloads: 23

American bibliography : a chronological dictionary of all books, pamphlets and periodical publications printed in the United States of America from the genesis of printing in 1639 down to and including the year 1820 ;with bibliographical and biographical notes (Volume 3) - **Evans**, Charles, 1850-1935
Includes index (v. 14)
Keywords: American literature; Printing -- United States History; Printers -- United States; United States -- Bibliography
Downloads: 15

American bibliography : a chronological dictionary of all books, pamphlets and periodical publications printed in the United States of America from the genesis of printing in 1639 down to and including the year 1820 ;with bibliographical and biographical notes (Volume 2) - **Evans**, Charles, 1850-1935
Includes index (v. 14)
Keywords: American literature; Printing -- United States History; Printers -- United States; United States -- Bibliography
Downloads: 9

American bibliography : a chronological dictionary of all books, pamphlets and periodical publications printed in the United States of America from the genesis of printing in 1639 down to and including the year 1820 ;with bibliographical and biographical notes (Volume 6) - **Evans**, Charles, 1850-1935
Includes index (v. 14)
Keywords: American literature; Printing -- United States History; Printers -- United States; United States -- Bibliography
Downloads: 16

Sort results by:
> Relevance
Average rating
Date
Date added

Group results by:
> Relevance
Mediatype
Collection

Refine your search
Collection
americana
cdl
File formats
MP3 files
Any audio file
Movie files
Images

Related collections

Figure 4.4: A SERP (search engine result) page from the Internet Archive. On the right hand side, the user can find advanced sorting and parametric filtering options. Source: `http://www.archive.org/search.php?query=evans+american+bibliography`.

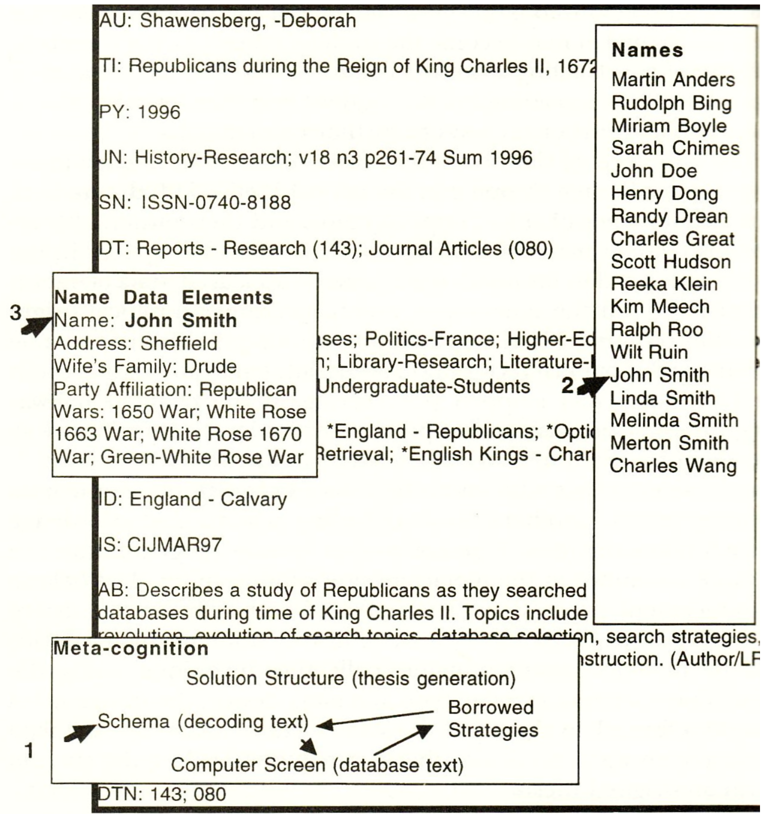
on the custom displayed by PhD-students in history to write down the names of authors and other people of importance to their research domain on index-cards. According to Cole, the purpose of the device was threefold:

> 'to (1) facilitate pattern recognition [. . .] based on the names-drawer schema; (2) show [. . .] the global structure or metacognition[25] of his or her own information processing; and (3) give the IR system details of the student's name-drawers schema so that it can use these data elements to formulate the student's query to the system's database (Cole, 2000a).

Figure 4.5 shows Cole's interface model based on his 'names-drawer' schema, consisting of four separate windows: The 'metacognition' window (the window numbered #1 by Cole, in figure 4.5), which acts as the starting point for the information-seeking process. This metacognition window represents the student's (internal) cognition process of mentally mapping the information; two 'microcognition' windows, one 'drawer' window, in which the individual items can be collected (in the form of names, window #2) and another (#3: 'Name data elements') in which the berries (the names) can be selected and added to the names-drawer (or berry basket). The fourth window (which is not explicitly given a number in figure 4.5) is the base window, where the bibliographic metadata is presented.

Cole's name-drawer schema provides an excellent example of how to apply observed behavior of historians into a conceptual model. It applies not only

[25] i.e. the student's mental contextual representation of the knowledge domain.



AU: Shawensberg, -Deborah

TI: Republicans during the Reign of King Charles II, 1672

PY: 1996

JN: History-Research; v18 n3 p261-74 Sum 1996

SN: ISSN-0740-8188

DT: Reports - Research (143); Journal Articles (080)

ıses; Politics-France; Higher-Ed
ı; Library-Research; Literature-I
Undergraduate-Students

*England - Republicans; *Opti
Retrieval; *English Kings - Char

ID: England - Calvary

IS: CIJMAR97

AB: Describes a study of Republicans as they searched
databases during time of King Charles II. Topics include
revolution, evolution of search topics, database selection, search strategies,
ıstruction. (Author/LF

DTN: 143; 080

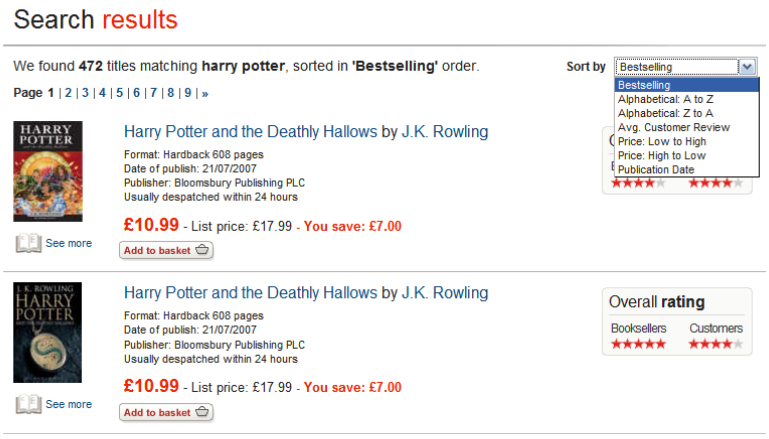
Names
Martin Anders
Rudolph Bing
Miriam Boyle
Sarah Chimes
John Doe
Henry Dong
Randy Drean
Charles Great
Scott Hudson
Reeka Klein
Kim Meech
Ralph Roo
Wilt Ruin
John Smith
Linda Smith
Melinda Smith
Merton Smith
Charles Wang

Name Data Elements
Name: **John Smith**
Address: Sheffield
Wife's Family: Drude
Party Affiliation: Republican
Wars: 1650 War; White Rose
1663 War; White Rose 1670
War; Green-White Rose War

Meta-cognition
Solution Structure (thesis generation)
Schema (decoding text)
Borrowed Strategies
Computer Screen (database text)

Figure 4.5: Cole's model, based on his 'names-drawer' schema: 'A proposed IR screen for Ph.D. history students accessing information via an electronic database. Student (1) selects "schema" in the metacognition window, which brings on screen the "names" window; (2) student selects "John Smith"; "name data elements" window for John Smith appears' (Cole, 2000a).

to history PhD-students, it can also be used to help ‘proper’ history scholars, as well as other groups of scholars that display similar information-seeking behavior. Although Cole never actually built his ‘name-drawer schema’, it does bear some striking similarities to the shopping cart (or basket) paradigm which is prevalent on the web, and well-documented as well (see figure 4.6 for an example).[26]



Search results

We found **472** titles matching **harry potter**, sorted in **'Bestselling'** order.

Sort by: Bestselling

Page 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | »

Harry Potter and the Deathly Hallows by J.K. Rowling
Format: Hardback 608 pages
Date of publish: 21/07/2007
Publisher: Bloomsbury Publishing PLC
Usually despatched within 24 hours

£10.99 - List price: £17.99 - **You save: £7.00**

Harry Potter and the Deathly Hallows by J.K. Rowling
Format: Hardback 608 pages
Date of publish: 21/07/2007
Publisher: Bloomsbury Publishing PLC
Usually despatched within 24 hours

£10.99 - List price: £17.99 - **You save: £7.00**

Overall **rating**
Booksellers ★★★★★
Customers ★★★★☆

Figure 4.6: A search result page for Waterstone’s on-line bookstore. Information objects (books), which are located through searching or browsing, can be added to a shopping basket. The shopping basket paradigm demonstrated on this page bears striking similarities to Cole (2000a)’s Name-collecting and Bates (1989)’ Berrypicking models. Source: `http://www.waterstones.com/waterstonesweb/simpleSearch.do?simpleSearchString=harry+potter`.

4.3 Information Visualization

> ‘Abstract information visualization has the power to reveal patterns, clusters, gaps or outliers, in statistical data, stock-market trades, computer directories, or document collections. (Shneiderman, 1996)

All throughot history, man has made use of visual representations. From the cave paintings found at Lascaux to medieval maps of the ‘known world’, from Da Vinci’s schematic drawings of the human body and flying machines, to family trees mapping centuries of family dynasties, all visualizations help people understand or grasp a concept in an instant. ‘A picture is worth a thousand words’, as the saying goes. Visualization, especially the visualization of physical information—has been around for ages, and has proven to be very effective.

[26] On a side not: Cole’s schema also bears some resemblance to Gert Pedersen’s browser for bibliographic information retrieval based on mathematical lattice theory (Pedersen, 1993). Lattices are mathematical sets composed of pairs of elements and a vector connecting these elements. Lattices in this context are used in a similar fashion as we will use semantic triples in this thesis (see also section 5.1.1): for computing and suggesting ‘similar’ items.

Scholarly Information Visualizations (IVs) have been around since the first scientist started recording his observations and experiments, in the form of diagrams and tables (Tufte, 1983).

4.3.1 Amplifying Cognition

In the previous sections we have shown that the design of a Berrypicking interface revolves around assisting the user in his information-seeking process. This can be achieved by supporting and enhancing the user's ability to recognize patterns and form a mental contextual schema of the research domain. Card et al. (1999) call this phenomenon 'amplifying cognition': the system should make use of 'external cognition' aids to augment 'internal cognition'.[27]

Information Visualizations[28] are ideally suited—according to Card et al. (1999)—to support and facilitate this internal sensemaking: 'Users can scan, recognize, and recall images rapidly, and can detect changes in size, color, shape, movement or texture' (Shneiderman, 1996). The purpose of Information Visualization—as part of this external cognition—is to amplify the user's cognitive powers, in this process of sensemaking, in six different ways (see table 4.1).

1) By increasing the memory and processing resources available to the users
2) By reducing search for information
3) By using visual representations to enhance the detection of patterns
4) By enabling perceptual inference operations
5) By using perceptual attention mechanisms for monitoring
6) By encoding information in a manipulable medium

Table 4.1: Visualization helps amplify cognition in six ways. Source: Card et al. (1999)

Information Visualization can be used in two ways: (1) within an electronic finding aid, as a data presentation and clarification aid—as an illustration—and (2) as a fully-fledged, self-contained navigation-and-presentation system.[29] As

[27] Cognition is best described as 'the intellectual processes in which information is obtained, transformed, stored, retrieved, and used' (Card, 2003). 'Internal' cognition is the internal sensemaking of the information and its context into a mental schema, whereas 'external' cognition is described by Card as 'the uses of the external world to accomplish some cognitive process' (Card, 2003).

[28] Card define Information Visualization as 'the use of computer-supported, interactive, visual representations of abstract data to amplify cognition' (Card et al., 1999).

[29] Proponents of information visualizations often take the interactive aspects of IV a bridge too far—in our humble opinion—and utilize them and rely on them as sole means of navigation through an infosphere. Often, with a website or an interactive environment, it is all too tempting to let technology take over—to let technology lead—and with this, along the way, the user-centeredness gives way. Effective visualizations should have users—*and* usage—as their

part of such an electronic finding aid interface, Information Visualizations can be seen as further transformations of the data tables—which we have seen in section 3.2.2—into 'visual structures'.

4.3.2 History Visualized: objects, dimensions, and mappings

> 'Visualizations [. . .] allow the historian to see patterns in the (visual) evidence, allowing for comparison across time, space, and scale' (Staley, 2003)

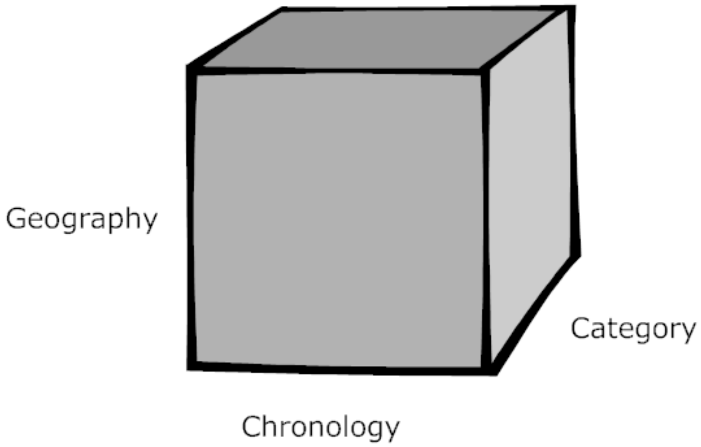


Figure 4.7: Douglas J. Cremer's "Cube of World History". Rather than a line, the cube organizes historical information as a three-dimensional space (Staley, 2003). A piece of historical information (primarily a document, but also events, or a person's lifespan) can be plotted against these three dimensions [x,y,z]: chronology (temporal), geography, category (topical). Adapted from: Staley (2003).

As we have seen earlier, historians find proof in—and only in—the form of historical primary resources. Trusted primary sources—usually in the form of documents—date from the actual time a historical event took place, and can therefore legitimately shed light on this event. In our semantic classification of this resource data we have identified several important objects which underly this historical evidence: *Documents* are *authored by people* (or groups of people: organizations)—*agents*—at a specific *place* and *time*. These documents are *published by agents* at a specific *place* and *time*, and deal with a specific *topic* or

main interest, and this can only be achieved by keeping clarity and simplicity in mind when developing and deploying (interactive) information visualizations. Using high-end, high-tech abstract visualizations—how appealing they may be to technologists and developers—may not fare well with users that set out to find information, who prefer text and 'typically scoff at a visual display as mere decoration' (Staley, 2003).

subject—which can be an *event*, another *agent*, or a more *abstract* topic (which can usually also be expressed in place and time).

We can restate this in an object-relation-oriented manner instead of a document-centered one: *history* revolves around *agents* (people and organizations), who take part in and instigate *events.* These events take place in *place* and *time*, all of which is evidenced by *documents*: all entities (documents, events, agents) can—according to Staley—be expressed visually as a function of scale (in relation to other entities), space and time.[30]

A slightly different—but also three-dimensional—approach to Staley's organization of history is brought forward by Douglas Cramer: any piece of historical information, and any scholarly research based on this information, can be seen as a point in a three-dimensional space, which is exemplified by what Cramer calls the 'Cube of World History' (see figure 4.7).[31] According to Cramer, every piece of historical information can be expressed in three dimensions: *chronology* (temporal), *geography* (geospatial), and *category* (topical) (via Staley (2003)).

4.3.3 Views: Visualization types

According to Shneiderman (1996) data can be visually mapped into in seven different view types: (1) 1-dimensional (textual documents, alphabetical lists), (2) 2-dimensional (maps,line graphs), (3) 3-dimensional (models of real-life objects, graphs with 3 axes), (4) temporal (timelines), (5) multi-dimensional (database visualization, and maps and graphs with additional filters and controls), (6) tree structures (hierarchies, directory and browsing structures), and (7) network (visualizations of (semantic) relationships and connections). Not all of them are suitable for use with historical data, and by historians, but some of these 'views' are well suited to enable historians to discover new items, new patterns, by amplifying external cognition.

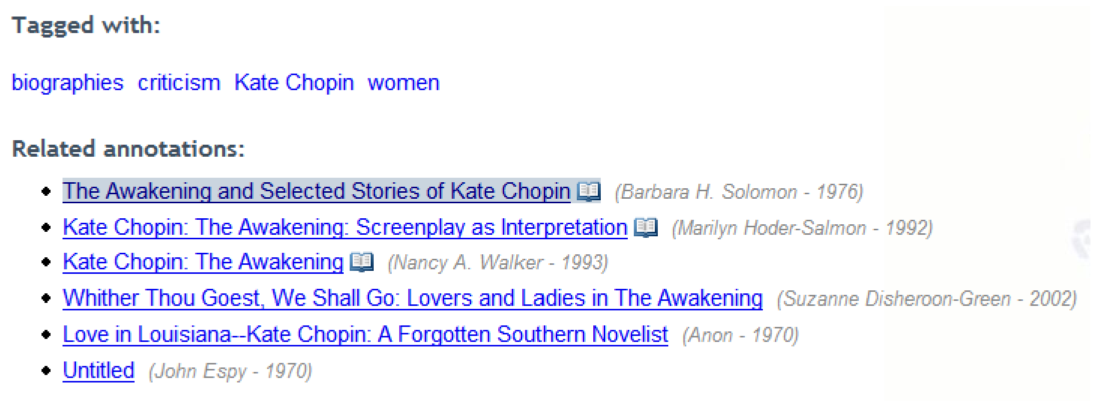
The Basis: Textual Views

The presentation of (collections of) information in (hyper-)textual format lies at the core of the World Wide Web. With the danger of stating the obvious, a

[30] These three 'dimensions' to history allow for historical data to be mapped into, in alignment with the IV reference model by Card et al. (1999).

[31] Please note that in its original conception, the 'Cube' was originally devised as a way to browse the products of historical research in a virtual three-dimensional environment, with each dimension subdivided into five parts. Each piece of historical research can therefore be placed on one of the nodes in the 5x5x5 grid. In my opinion, it is a visualization that has its weaknesses. Dimensions like time and location are quantitative in nature, and are easy to express in a spatial substrate. However, it is hard to quantify a 'dimension' as nominal in nature as A three-dimensional grid where two dimensions are quantitative and the third is not, is an impractical and artificial construction to say the least. The 5x5x5 subdivision seems arbitrary as well. As a metaphor, though, it remains powerful.

textual representation is essential in the development of an effective visualization of a collection of data on the Web. And, as shown in the previous chapter, historians are mainly text-oriented: most of their sources are in (running) textual form, so a textual representation cannot be overlooked. Furthermore, a hypertextual representation serves as the base foundation of every usable and accessible website. Furthermore, 'historians typically scoff at a visual display as mere decoration' (Staley, 2003), which makes them of secondary importance to historians. Whereas all other visualizations may be optional in an on-line finding aid, a (hyper)textual view is not.



Tagged with:

biographies criticism Kate Chopin women

Related annotations:

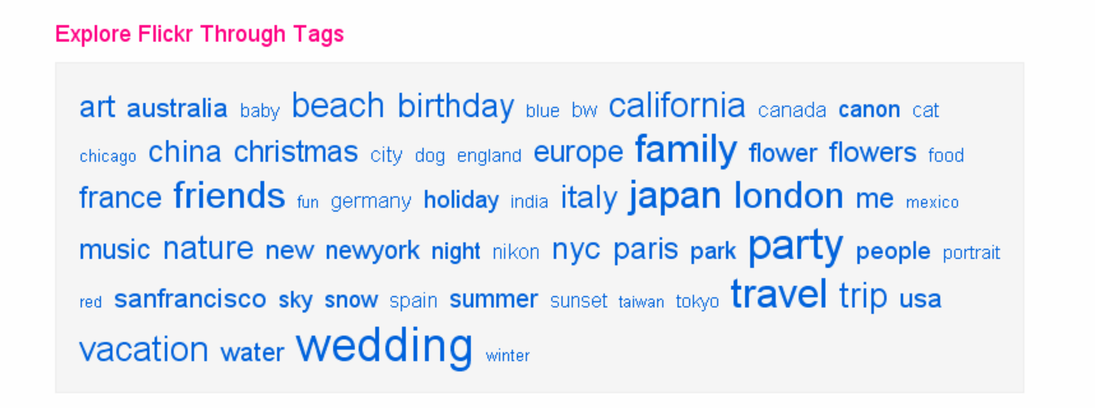
- The Awakening and Selected Stories of Kate Chopin 📖 *(Barbara H. Solomon - 1976)*
- Kate Chopin: The Awakening: Screenplay as Interpretation 📖 *(Marilyn Hoder-Salmon - 1992)*
- Kate Chopin: The Awakening 📖 *(Nancy A. Walker - 1993)*
- Whither Thou Goest, We Shall Go: Lovers and Ladies in The Awakening *(Suzanne Disheroon-Green - 2002)*
- Love in Louisiana--Kate Chopin: A Forgotten Southern Novelist *(Anon - 1970)*
- Untitled *(John Espy - 1970)*

Figure 4.8: An ordered list can be used to point the user toward other similar, and potentially interesting items, as seen on an annotation page for the book *Kate Chopin: A Critical Biography*. Source: SSSL (Society for the Study of Southern Literature) Bibliography (`http://www.missq.msstate.edu/sssl/betatest/?page=annotation&id=1058`).

Text can serve serves multiple purposes in a website: it can appear as running text, to serve as the page's content. Text can also provide a navigational structure, in the form of a (semi-)structured list. But text is most effective when used in a structured format (Nielsen, 2003). The most common and single most effective textual information visualization is an *ordered list*, especially when the order is used to emphasize the relative importance of an item (frequency, or 'relevancy') (Halvey and Keane, 2007; Rivadeneira et al., 2007). The search engine result page of figure 4.6 is one example of a list ordered by 'relevancy' (in this case the relative match of a book to a IR keyword query). Figure 4.8 provides another example of a ranked list: here a list of five similar (an potentially relevant) items is shown on an imprint's detail page in an on-line bibliography.

Nearly every website uses a frequency-ranked list to present important items to users. However, a growing number of websites are using another form of textual visualization: Tag clouds are rapidly gaining popularity, and are proven to be very effective when used in combination with ordered lists (Rivadeneira et al., 2007; Halvey and Keane, 2007; Hasan-Montero and Herrero-Solana, 2006). Tag clouds are 'visual representations of a set of words, typically a set of tags[32], in

[32] Tags are user-generated topical metadata. They are one of the most prominent features of



Explore Flickr Through Tags

art australia baby beach birthday blue bw california canada canon cat chicago china christmas city dog england europe family flower flowers food france friends fun germany holiday india italy japan london me mexico music nature new newyork night nikon nyc paris park party people portrait red sanfrancisco sky snow spain summer sunset taiwan tokyo travel trip usa vacation water wedding winter

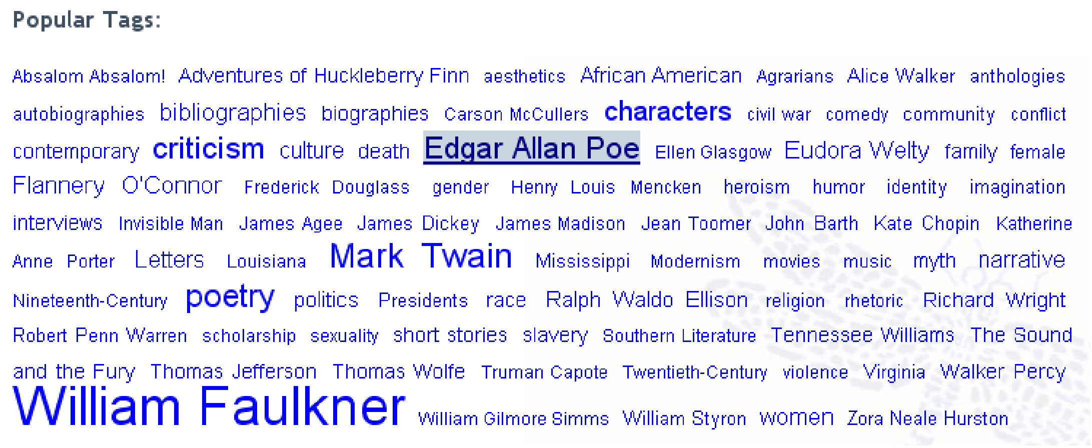
Figure 4.9: A tag cloud provides an instant overview of popular topics, on Web 2.0 photo sharing site Flickr.com. Source: `http://www.flickr.com/explore/`

which attributes of the text, such as size, weight or color can be used to represent features (e.g. frequency) of the associated terms' (Halvey and Keane, 2007). These clouds are ordered alphabetically and span a number of horizontal lines of text. Comparative studies have shown them to be slightly less effective than frequency-ordered (vertical) lists (Rivadeneira et al., 2007; Halvey and Keane, 2007), but they do appear to play a very important role in the information-seeking process; tag clouds support the user with 'impression formation' or 'gisting': a user can get a general overview of the underlying data set or and impression of the entities associated with the data by scanning the tag cloud (Rivadeneira et al., 2007). The alphabetical organization of a tag cloud can aid users to locate specific information easily and quickly, whereas the difference in font size is very important for the users to develop a quick sense of overview (Halvey and Keane, 2007). Figures 4.9 and 4.10 show two examples of such tag clouds, in both a 'social web' context (Flickr) and a document collection context (SSSL Bibliography).

Geographical: Maps

Geography is—along with *chronology*—the most important dimension to visualize history, and historical studies often make use of geographical visualizations: 'Historians accept maps and atlases as visual secondary sources more readily than any other type' (Staley, 2003). As historical events take place in place and time, the geographical dimension lends itself perfectly for creating effective information visualizations. Computer-based Geographical Information Systems (GIS) have been in use by scholars and professionals, and continue to be very popular among various user groups. The implementation of mapping solutions

the 'social web', together with similarity lists and social bookmarking. On a community-wide scale, the aggregated metadata form so-called folksonomies: people-generated taxonomies. Tag clouds can also be used to represent other objects, like ontology-based metadata, but the term 'tag cloud' is commonly used to denote all types of 'clouds'



Popular Tags:

Absalom Absalom! Adventures of Huckleberry Finn aesthetics African American Agrarians Alice Walker anthologies autobiographies bibliographies biographies Carson McCullers characters civil war comedy community conflict contemporary criticism culture death Edgar Allan Poe Ellen Glasgow Eudora Welty family female Flannery O'Connor Frederick Douglass gender Henry Louis Mencken heroism humor identity imagination interviews Invisible Man James Agee James Dickey James Madison Jean Toomer John Barth Kate Chopin Katherine Anne Porter Letters Louisiana Mark Twain Mississippi Modernism movies music myth narrative Nineteenth-Century poetry politics Presidents race Ralph Waldo Ellison religion rhetoric Richard Wright Robert Penn Warren scholarship sexuality short stories slavery Southern Literature Tennessee Williams The Sound and the Fury Thomas Jefferson Thomas Wolfe Truman Capote Twentieth-Century violence Virginia Walker Percy William Faulkner William Gilmore Simms William Styron women Zora Neale Hurston

Figure 4.10: A 'tag' or subject cloud as it appears on the overview page of the newly proposed overview page of the SSSL's on-line bibliography. Source: `http://www.missq.msstate.edu/sssl/betatest/`.

in a web-based environment is a more recent development (Huynh et al., 2007).

As modern on-line maps become more and more manipulable, and easier to implement from a technical standpoint, the number of web sites making use of these visualizations as part of their interfaces is growing rapidly, and as a result, more and more people are becoming familiar and accustomed to using them. These on-line maps have become the perfect example of Shneiderman's visual information-seeking mantra: 'Overview first, zoom & filter, details-on-demand' Shneiderman (1996). Historical information can be plotted on a map by transforming nominal location information into a set of quantifiable geographical coordinates. This yields a basic two-dimensional data-structure that can be plotted on a map. Including additional controls, such as faceted filtering and grouping options and additional visual marks as shapes and colors to encode extra variables such as frequency or to distinguish various classifications/types (as genre), can transform the two-dimensional map into a multi-dimensional, multi-variable interactive structure (Card et al., 1999; Shneiderman, 1996). Figure 4.11 shows an implementation of such an interactive map, with the aforementioned imprints on abolition plotted on the map by city of publication. It must be noted, that using modern maps for visualizing historical—non-modern—data does pose a danger. It is of utmost importance to note that when aggregating and transforming historical data the very real danger of creating anachronisms exists.[33]

[33] This especially poses a danger when using new mapping technology to display century-old data. Current maps show present-day states (Mass., Penn., NY) and place names (like Washington, DC) in the 17th and 18th century, whereas the state of Massachusetts and the national entity of the 'United States' did not come into existence until much later. Using modern satellite imagery, and allowing users to zoom in on demand (Shneiderman, 1996) makes matters even worse. Most mapping packages offer a solution in the form of using overlays of historical maps and of developing layers to show geographical entities as they existed at the desired moment in time.

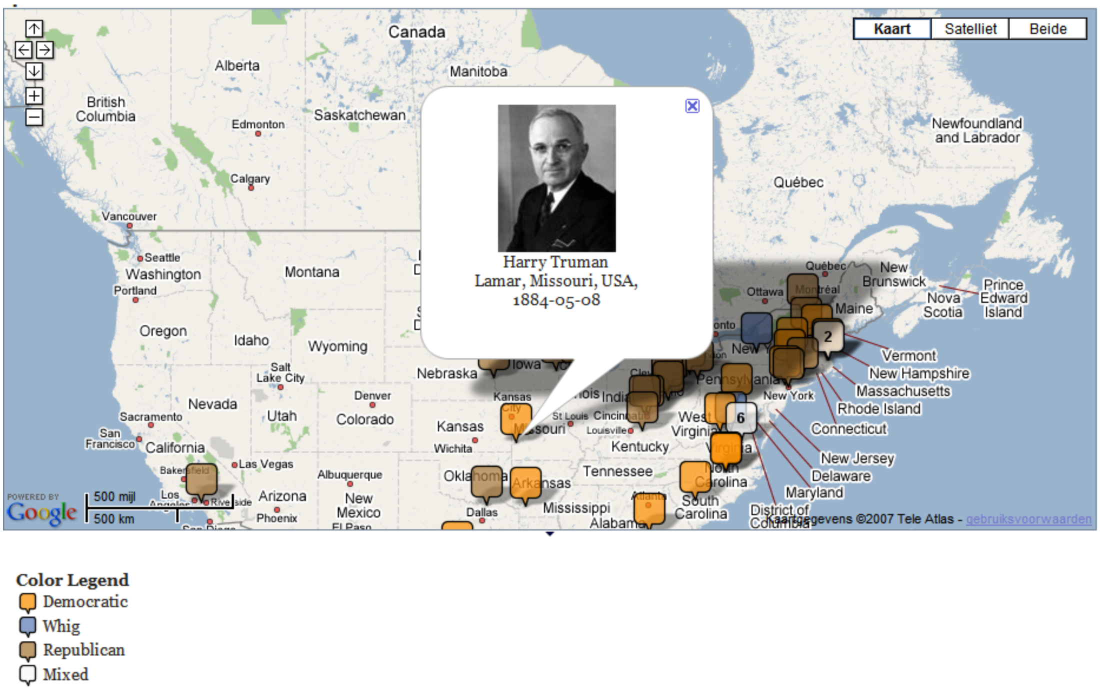


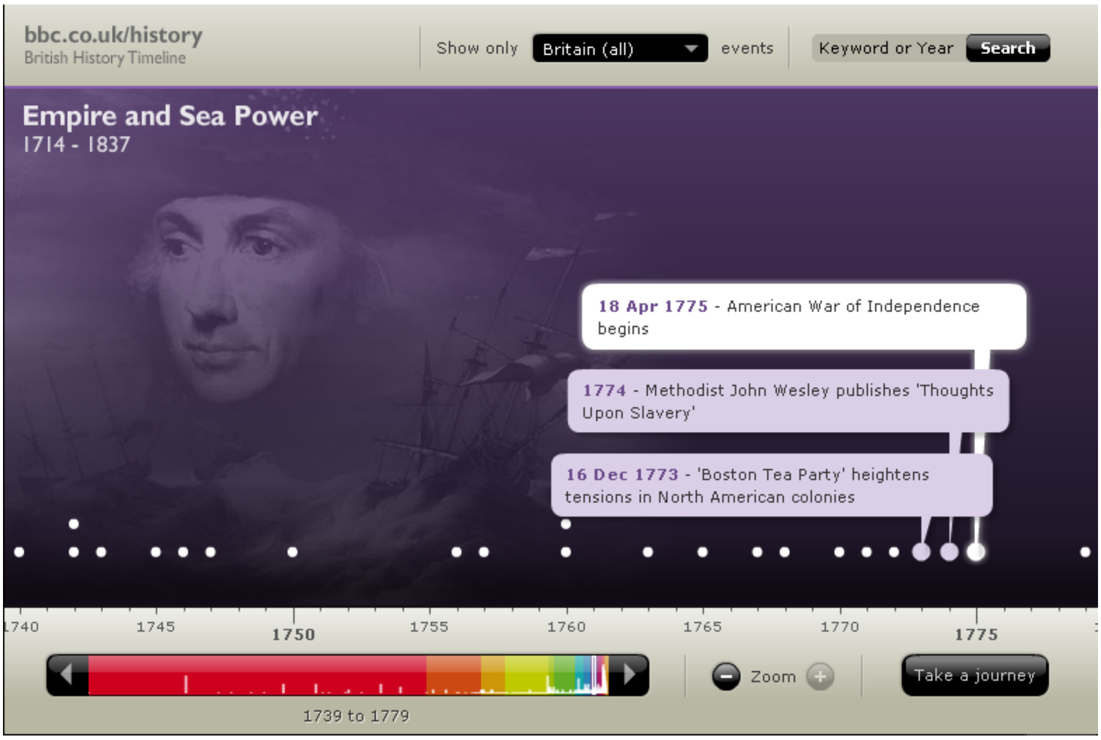
Figure 4.11: The Google Map API can also be used to plot historical information. Source: *Simile Exhibit: Presidents* `http://simile.mit.edu/exhibit/examples/presidents/presidents.html`.

Temporal: Timelines, Time Series Charts

Evans' initial organization of his *American Bibliography* into a chronological arrangement underlines the importance of temporal context to historical research. Every piece of historical information can be—directly or indirectly—linked to a date or a time interval. Nearly all entities in the *Evans* corpus have a temporal property associated with them. The prominence of the temporal dimension in historical data makes the timeline an excellent visualization, one that is very familiar to historians. According to Staley (2003), a timeline is 'the most elementary diagram used in history', fundamental in the way that its simplicity can help students and school kids alike to get their events straight, but also in the way that it fits in closely with the way we think of the nature of time and history, a continually flowing sequence of events, arranged chronologically.

A timeline places individual objects in a temporal dimension. A timeline is—in most cases—one-dimensional: the temporal dimension is usually arranged horizontally, with individual objects—events, documents, people's lives—placed on this line. The vertical axis—if present—is used solely to distinguish between different categories or subdivisions of data. Figure 4.12 shows a prime example of the use of an interactive timeline in a historical context.

In order to the visualize aggregated data in a temporal dimension a time series chart (or line graph) is more suited. A line graph allows for quantitative variables



Show only Britain (all) events

18 Apr 1775 - American War of Independence begins

1774 - Methodist John Wesley publishes 'Thoughts Upon Slavery'

16 Dec 1773 - 'Boston Tea Party' heightens tensions in North American colonies

Figure 4.12: An interactive timeline on the *BBC* History website allows for interactive manipulation and all four actions of Shneiderman's visual information-seeking mantra ('Overview first, zoom & filter, details-on-demand') Shneiderman (1996). Source: *BBC British History Timeline* (`http://www.bbc.co.uk/history/interactive/timelines/british/index.shtml`).

to be displayed, and to be juxtaposed against time. A line graph gives the user or reader a quick overview of the subject(s) at hand, and how its frequencies changed over time. Figure 4.13 shows the total number of imprints published per year in early America, based on the *Evans* data set, giving a quick impression of the growth in printing output from 1640 to 1800, showing among other things, a decline during the Revolutionary War (1775-1783).

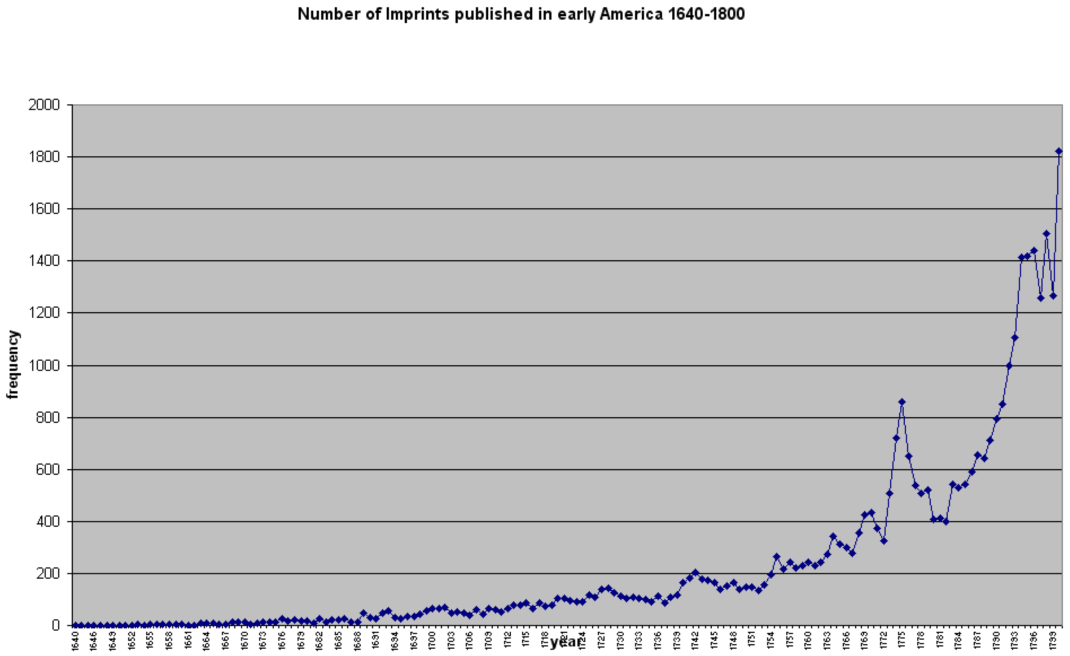


Figure 4.13: Total printing output per year, 1640-1800. This line graph shows a steady increase in publications, with a significant decline during the years of the Revolutionary War (1775-1783). Source: *Evans* Corpus/SWHi Ontology.

Contextual: Tree and Network Graphs

In historical practice, the fourth common form of visualizations is that of diagrams showing relations between objects. Scientists have been using diagrams for centuries, trying to classify and conceptualize phenomena (Card, 2003). One cannot imagine historical visualizations without thinking of family trees (or causality diagrams). These tree or network diagrams are often used to denote relations between objects. They can be used to visualize the causality between related events events, or to chart the—hierarchical—structure of a society or of a family dynasty (as is shown in figure 4.14). These diagrams can express with a few lines and arrows, within a few square inches of paper, intricate relations that cannot be described with a mere few words.

In a Semantic environment, which is based on object entities and the relations between them, a network graph is perhaps the most obvious of candidates for

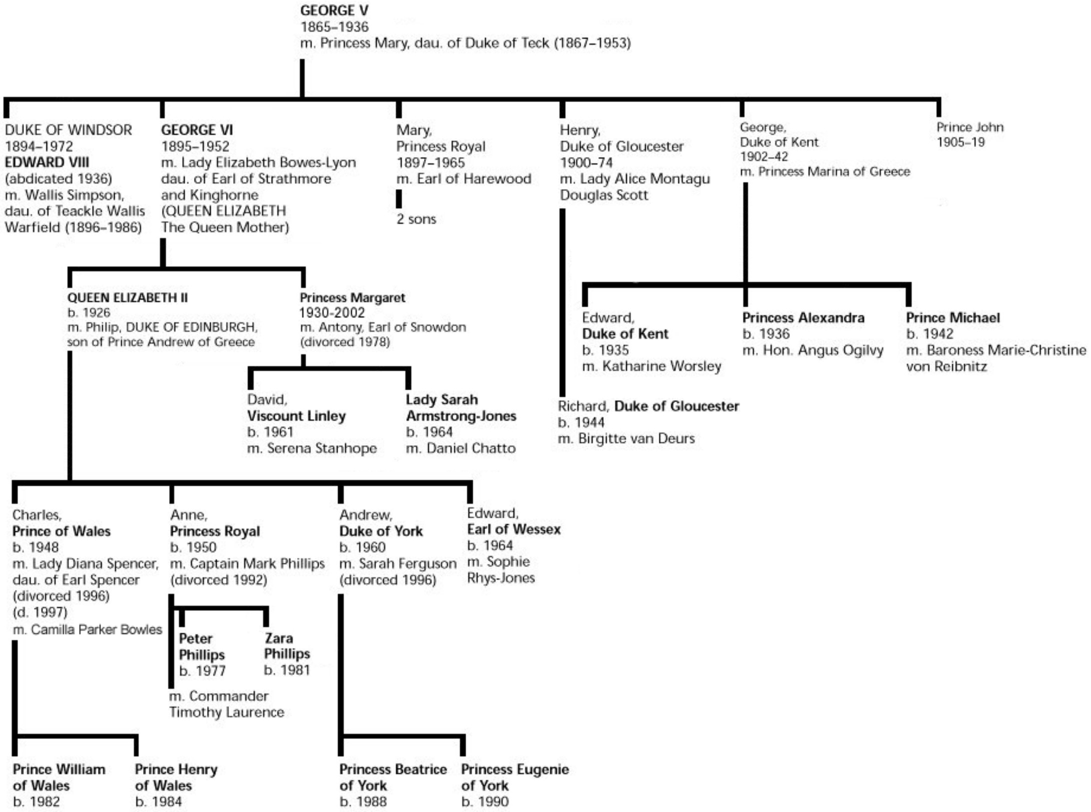


Figure 4.14: A tree diagram, shows the lineage of the Windsor's, the British Royal Family. Source: *Infoplease.com: Britain's Royal Family Tree* (`http://www.infoplease.com/spot/royal3.html`).

visualizing such intricate relations. Network graphs such as TouchGraph[34] and Clustermap[35] are commonly used to visualize ontologies (Geroimenko and Chen, 2002; Fluit et al., 2002; Kimani et al., 2002; Fahmi et al., 2007). Figure 4.15 shows a Touchgraph visualization of the actor John Cusack, and his network of films and co-stars. With a collection of over two million triples, the SWHi ontology has a plethora of meaningful relations in store in order for historians to discover and explore, and network graphs might be a good way to visualize these relations. Network graphs can also help to assist the 'context-building' which takes place during the information-seeking process by providing an external cognitive aid visualizing the items within that context in relation to one another.

Although potentially effective as a visualization, the type of interpersonal relations currently available in the SWHi ontology do not include explicit family or other hierarchical relations to allow for tree visualizations such as the one in figure 4.14. This will perhaps be the case in a future version of the ontology, enriched with data and relations from other sources. However, (hyper)textual tree and network structures are applicable all over the web, and can be used throughout the *Evans* infosphere and information workspace, when implementing the browsing architecture and framework based on the taxonomies and classifications described in this (and the previous) chapter.

4.4 Chapter Summary

In this chapter we have attempted to answer the following question: 'what should a search interface for use by historians look like?' We started out with a set of requirements, which were based on both the user and the data. We have attempted to investigate these findings in the light of potential interface elements for an on-line berrypicking electronic finding aid.

An ideal interface for an on-line electronic finding aid is one that is only be usable for historians, but one that is also useful. Such an interface should allow the user unrestricted access to the data in the way that fits best with his information-seeking habits. It should furthermore allow for multiple—and advanced—ways to order, transform and view the data. Information Visualizations not only accommodate this behavior, but also augment and empower the user's ability to seek, find, and discover interesting items from within the dense information forest of the *Evans* collection, as we have attempted to demonstrate.

[34] http://www.touchgraph.com/.

[35] http://www.techmap.ca/.

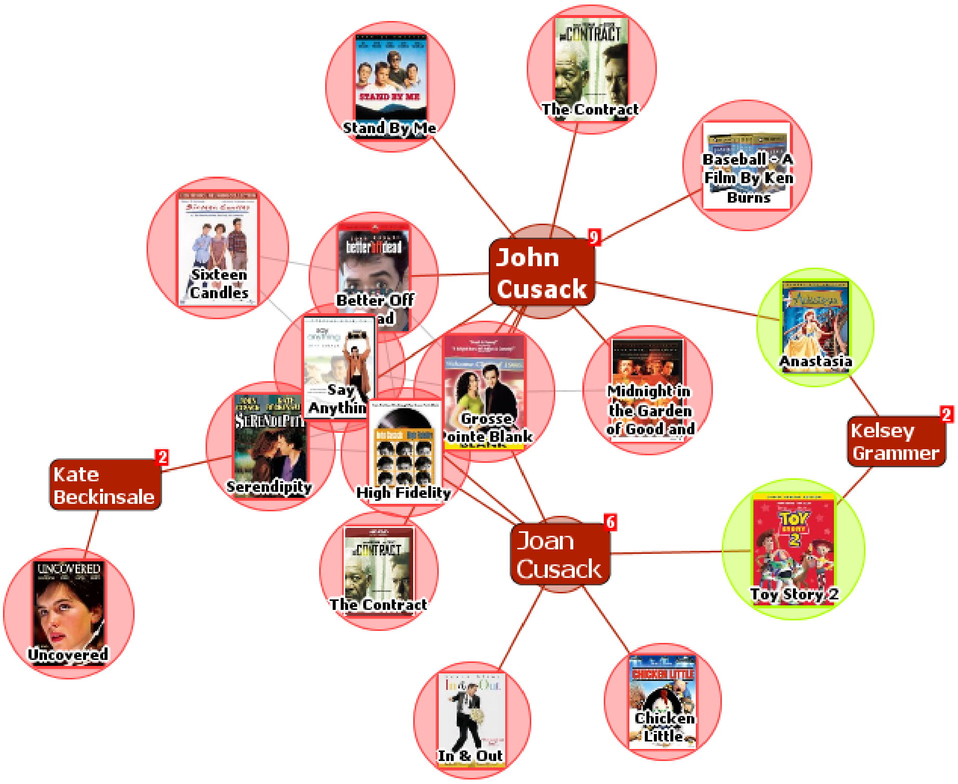


Figure 4.15: Touchgraph network graph for the actor John Cusack, showing the actor in relation to other movie stars, via movies they co-starred in. Source: Touchgraph Amazon Browser (`http://www.touchgraph.com/TGAmazonBrowser.php?query_type=Movie&keywords=John+Cusack`).

Chapter 5

Semantically-Enhanced Berrypicking: A New Proposed Model

The goal of this thesis is to arrive at a set of requirements for a historical 'berrypicking' interface, and to formulate a model based hereon. For this reason, this chapter does not deal with the implementation of these requirements and recommendations into a fully working definitive version of the SWHi On-line Finding Aid to Early American Imprints. It rather deals with the formation and the proposal of a new model and an algorithm for information-seeking within an electronic finding aid environment. In this chapter, we investigate the power an organization of the data in semantic triples can bring such a finding aid environment. We try to combine this power and the precision of the Semantic Web (aka Web 3.0) with the flexibility and the usability of Web 2.0 (Social Web) interfaces, in order to form a new model: a Berrypicking Interface Model for Historical Data, which we call 'The Semantically-enhanced Berry Basket'. This model will be illustrated and explained with various examples.

5.1 Getting the Best of Both Worlds: Combining the Social with the Semantic Web

In recent years, many articles have been written on how to combine the two competing visions of a 'future web' into one single vision. These articles have dealt with either adding more precise semantics to Social applications[1], or the exact opposite: use the Social Web (community of users, extracting semantics from folksonomies) to enhance ontologies and Semantic applications (Ankolekar

[1] For instance by embedding Semantic code or tags to existing Webpages. See also Khare and Çelik (2006).

et al., 2007; Herzog et al., 2007; Khare and Çelik, 2006). Although the two 'future webs' appear to follow two diverging paths to the future, they can indeed be complementary and can draw from each others strengths (Ankolekar et al., 2007). In order for the idea of the Semantic web to take hold, a usable, and user-centered approach is required (Ankolekar et al., 2007; Huynh et al., 2007; Finin et al., 2005). We fully endorse this attempt to harmonize both visions.

However, to us, the most interesting prospect of bringing features of the Social to the Semantic web does not lie in its social, communal, or networked features, but in its interactive interface and information-augmentation features. The Social Web, with its usable, flexible and user-centered interfaces, can give the social web the 'killer app' it requires. (Herzog et al., 2007; Huynh et al., 2007; Finin et al., 2005; Fahmi et al., 2007). Although Social and Semantic applications use very diverse data to power their interfaces, we think that we can augment the Semantic web experience by borrowing interface elements from Social applications. Data is data, regardless of whether it was obtained via user input, analysis of visitor statistics and website usage, or from a large (historical) information corpus. The Information Visualizations we studied in the previous chapter—subject clouds, interactive maps and timelines are prime examples of interactive Web 2.0 visualizations, are therefore ideally suited to enhance the information-seeking experience of the scholarly users of electronic finding aids.

5.1.1 Semantic Similarity Suggestion: Setting Metadata to Work

Throughout in this thesis we have emphasized the information-seeking goal of historians to seek potentially relevant pieces of information. We have also hinted[2] at the potentially powerful application of semantic triples to provide and identify these pieces of information. It is perhaps time that we combine the two and elaborate on what we think of as one of the unsuspected strength of a Semantic data structure, an ontology. But first, let us paint you an interesting picture.

How amazing would it be if it was possible that, when browsing for information in an on-line archive, or when trying to locate books for your research library's online catalog, have a book or an article suggested to you, purely based on the page you are looking at right now, or the articles you have set aside for downloading and printing?[3] These suggestions might very well be books and articles that you might not have found this easily on your own, perhaps because the its title does not contain the specific keyword you searched for, or because it was written by an

[2] In section 3.2.2, to be exact.

[3] We realize that automatic suggestion and providing users with a list of potentially interesting items is used in many commercial and social on-line enviroments. Its use in a scholarly environment, however, still remains largely unexplored, especially with bibliographic metadata and ontologies.

author you would not have thought of initially. The reliance of current electronic finding on keywords carries with it a lot of implicit dangers: it can exclude works in different languages or from different eras, when people used different words to describe a certain phenomenon. In Information Retrieval terms this shows that in current IR systems recall is not maximized: not all desired items are returned by the keyword search. The opposite can also be the case: one word can have different meanings in different contexts. Searching for an ambiguous keyword can return a lot of irrelevant items. In this case, precision is negatively influenced, if we stay in IR terms.

Our users, the historians, have already adapted their information-seeking behavior to address this phenomenon. Historians have compensated for the impresision and ambiguity of keywords by employing different information-seeking techniques, like name-collecting (Beattie, 1990; Orbach, 1991; Duff and Johnson, 2002).[4] Historians also use proper names for a different purpose: they use names to form a mental overview of their research subject, and with this 'information in their heads' they discover relations to other pieces of information, to other works, to other objects, which are similar or at least related to the items they already know of. Implicitly, they were doing what could have potentially be achieved automatically all this time: they relate objects with shared characteristics to one and another.

The Semantic Web can cater to this need. Although mostly intended for use by and between automatic agents, the entity-relationship-based setup of semantic triples is ideally suited to help guide information-seekers by pointing out interesting relations, and even better still, of similar entities based on these relations.

Pointing out collaborating authors based on a certain book or article is easy. Pointing out a different work of an author is also not a hard task: users can usually do this without the help of a high-tech electronic finding aid. They do not even need an elaborate solution such as the Semantic Web to accomplish such a thing: a simple metadata structure can do the trick, and linking an author to a work is a simple matter of creating a hyperlink. But when elaborate, inferred relations between and across objects are needed, a semantic structure based on triples is much more suited to perform these tasks.

But even then, all of this can be done with a traditional database, albeit with some difficulty and some restrictions. In the relational model, object entities are stored in separate, 'normalized' tables, and connections between tables are sparse. With these sparse links, object entities can be related to other entities, and we can infer a basic form of 'similarity' based on even a sole relation. In a (semantic) triple structure, every 'record' instance is a triple, and denotes a relation between two items. These items can be of the same entity type, but they can also be of very different entity types. The next 'record' instance—the next triple—can consist of different entity types alltogether. Even the semantic

[4] See also section 2.2.2.

relation between the two items—the predicate—can denote an entirely different relationship between the subject and the object. The flexibility and simplicity—stored within this simple yet descriptive data structure—allows for powerful and meaningful inferencing between all entity instances in the ontology (regardless of entity type). The separation and the relatively sparse and 'meaningless' relations of a relational tabular structure make it per definition less capable of computing and inferring shared relations than if the same data was stored in a triple, in an ontology.

	Subject	*Predicate*	*Object*
a.	Book 1	hasAuthor	Person 1
b.	Book 1	hasSubject	Topic 1
c.	Book 1	hasSubject	Topic 2
d.	Book 2	hasSubject	Topic 1
e.	Book 2	hasSubject	Topic 2
f.	Person 1	likes	Person 2
g.	Person 2	hasRead	Book 1
h.	Person 2	hasInterest	Topic 1
i.	Person 2	hasInterest	Topic 2

Table 5.1: Nine example triples. Sharing a common link to an object is an indicator of similarity. See also figure 5.1.

For the purposes of this thesis, we define that two entities can be considered *similar* (or related)[5] if they have *at least two semantic connections* via triples to other entities in common *of which they both are the subject.*[6] The more predicates and objects they 'share', the more similar they are.[7] Perhaps it is best to illustrate the concept of this multi-dimensional semantic similarity with some (fictional) triples as an example (See table 5.1 and figure 5.1). Table 5.1 and figure 5.1 show six entities—two of each entity types—that are interconnected through a total of

[5] On the web, 'similar' and 'related' are used interchangeably in this context. Please note that there *is* a significate between these two terms and the term 'relevant': Relevancy is in the eye of the beholder. Only a user can decide, for himself, whether a similar object are in fact relevant.

[6] The reason why we have not chosen one common triple as the minimum threshold is because all objects of the same type automatically have one common triple: if A is of the type Topic, and B is also of the type Topic, this does not automatically mean that these two Topics are similar. Furthermore: some items in the ontology only have one triple which they are the subject of: most Topics in the SWHi ontology, although they may be the *object* in many triples, are only the *subject* in one triple: again, that of rdfs:Type.

[7] One can ask the question: What does this mean the other way around? Does a commonly shared relation of two entities to a third entity via a triple, in which they are both *objects* also not count towards a higher similarity? Not as far we are concerned; This would mean that just because a person has read a book and likes a certain kind of ice cream, these four entities are similar as one another. In certain cases, a reversed common relation many denote some form of similarity, but it depends on that specific case.

nine triples. Book 1 can be considered similar to Book 2, as they have two triple pairs in common ([b,d], to Topic 1; and [c,e], to Topic 2). Furthermore, Person 2 can also be considered similar to Book 2, as they share triple pairs [d,h] and [e,i]. In this example, Book 1 can also be considered similar to Person 2 (through pairs [b,h] and [c,i]).[8]

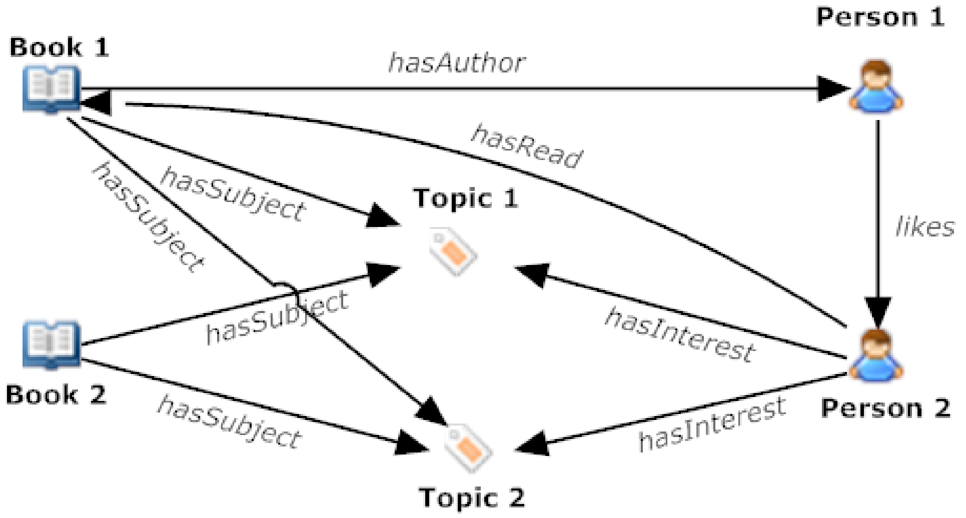


Figure 5.1: An example semantic network, consisting of nine triples. Book 1, Book 2 and Person 2 can be considered mutually 'similar': they have two triple-connections in common (they all share triples with Topics 1 and 2 as objects).

5.1.2 Further enhancements to the Suggestion Mechanism

The formula/algorithm of computing similarity between items can be surprisingly powerful, even in its current rather crude form. At this time the system does not take into account *what kind* of relation exists between two object, and how strong (or trustworthy) this relation it. In a later stage of the SWHi project, a weighting mechanism can be applied to make a clear distinction between predicates denoting 'strong' and predicates denoting 'weak' relations between entities. We may even decide to let the user apply and manipulate these weights himself: a print historian might put a higher value and trust in provenance-based relations more than a scholar interested in social history, who might want to give a higher weight to predicates that denote interpersonal and other social relations (friend, acquintance, sister, co-author etc.).

[8] Please note that in a real ontology, Book 1 would be considered more similar to Book 2 than to Person 2, as they would additionally share one further predicate and object [Book1, hasType, Document],[Book2, hasType, Document]; every entity also has a type.

Furthermore, the ontology—which forms the basis of this suggestion mechanism—does not have to remain restricted to the current bibliographic metadata. Other data sources can also be added to further include biographical or event data, or bibliographic metadata from alternative sources. The more data sources, the more triples we can use to compute similarity, and the stronger the statistical basis becomes for making automatic suggestions. The power lies in the numbers in this case.

Social and user-generated data can also form a potential additional data source with which further similarity can be computed. Every web server automatically accumulates visitor data, which can be used towards this goal. Social data like tagged and socially bookmarked items can also used to cluster and relate entities to one another. However, when merging these data sources with the existing set of—quality-controlled—authoritative data, special attention should be paid with respect to safeguarding and enforcing the quality of the data set. Additional appropriate weights should be applied to distinguish between the original data and the additional data sources.

5.1.3 Other potential applications for Similarity Suggestion

The semantical similarity mechanism, as it is described above, is not restricted to a single implementation with a historical document collection. It can, in theory, be applied to every well-structured set of data. It has a plethora of possible applications in a wide variety of (scholarly, bibliographic, as well as many other) fields. It can be implemented to work with every environment in which a desire or need for pattern recognition or item suggestion exists. This can be in a library catalog, an online shopping environment, or a Social Web application. This mechanism can be used to identify missing connections or ‘missing links’ in any data set. One potentially interesting application with predicting such ‘missing links’ can lie in suggesting a list of likely candidates for issues involving author attribution.

5.2 A Berrypicking Interface Model for Historical Data: The Semantically-enhanced Berry Basket

Charles Cole’s Name-Drawer schema for Berrypicking and Name-Collecting provides us with an excellent basis toward developing an on-line electronic finding aid. With a few alterations and several augmentations we can apply this name-drawer/shopping basket model to suit our purposes. These augmentations come from both the Social Web (assisting the user by displaying Interactive Information Visualizations; see figures 5.2 through 5.) and the Semantic Web (by offering inferred similarity suggestions described in the previous section). But let us first discuss the diffence with Cole's Name-Drawer schema:

The first modification to Cole's schema we would like to propose for an online application of this model is that it can also be applied to other entities than merely names. Although we do recognize that name-collecting forms an integral part of the historian's IB, other objects should also be added to the basket: like with an on-line bookstore (see figure 4.6) the books (imprints) themselves may also be added to the berry-shopping basket—and any other object type for that matter. Furthermore, we can do away with all the separate windows, as current web *usance* (usability: no 'frames' or 'popup windows') and technologies (AJAX[9], Rich Internet Applications) dictate that all interface actions should take place in a single browser window. Also, the metacognition window does not need to be visualized implicitly in a berrypicking interface, as in our opinion the cognitive actions are performed subcontiously and implicitly by the users. The metacognition process of 'knowledge crystallization' does not to be visualized implicitly as part of the interface itself.

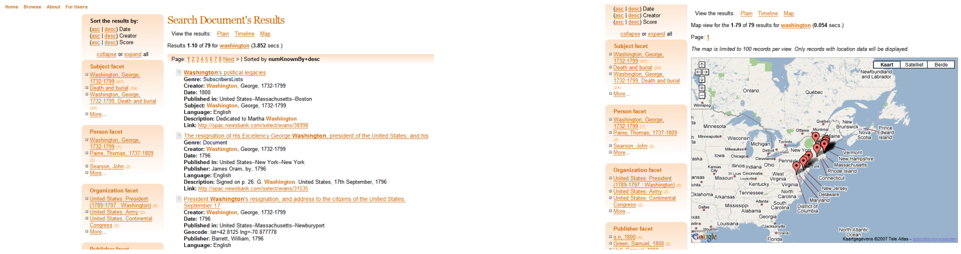
Finally, we propose an extension to this model in a manner that the 'berries' are not merely stored in a basket or a container, leaving the hard work—the mental contextualization—to the historian, as is the case with Cole's schema. Instead the berries that berries that are collected by the users—along with the (semantic) metadata properties they share between them—are used by the system to automatically suggest other similar, and potentially 'relevant' objects. These suggested objects are selected on the basis of the percentage of properties and semantic relations they have in common with the objects in the basket (see the previous section). We would like to call our implementation of—and addition to—Bates' Berrypicking model and Cole's Name-Drawer Schema the 'Semantically-enhanced Berry Basket'.

The algorithm that goes along with this model is as follows:

1. *The user goes looking for interesting pieces of information in the collection*, information which might help him in his research. [10] At any page, on every level of the hierarchy, potentially interesting items (berries) are brought to his attention, both textually (via frequency-ranked lists (figure 5.4), lists of 'similar' items (figure 5.5), and tag clouds (figure 5.6)) and graphically (through the option of displaying any list of objects in several views: Maps (figure 5.2, right), Timelines (figure 5.3, left), Network (figure 5.3, right))

[9] AJAX stands for *Asynchronous JavaScript and XML*, a technology that can be used to create fast-responding interactive web applications. User actions trigger Javascript 'events', which allow for packets of information (formatted in XML) to be sent asynchronously, i.e. without the need to load a new web page or reload the entire web page that on which the events were initiated.

[10] This can be done by searching (see figure 5.2, left), browsing, any method of information-seeking/berrypicking as described in the previous sections.



Sort the results by:
(asc | desc) Date
(asc | desc) Creator
(asc | desc) Score

collapse or expand all

Subject facet
- Washington, George, 1732-1799 (37)
- Death and burial (24)
- Washington, George, 1732-1799, Death and burial (24)
- More...

Person facet
- Washington, George, 1732-1799 (7)
- Paine, Thomas, 1737-1809 (2)
- Searson, John (2)
- More...

Organization facet
- United States. President (1789-1797 : Washington) (5)
- United States. Army (2)
- United States. Continental Congress (2)
- More...

Publisher facet

Search Document's Results

View the results: Plain Timeline Map

Results **1-10** of **79** for washington (**3.852** secs.)

Page: 1 2 3 4 5 6 7 8 Next > | Sorted by **numKnownBy+desc**

Washington's political legacies
Genre: SubscribersLists
Creator: Washington, George, 1732-1799
Date: 1800
Published in: United States--Massachusetts--Boston
Subject: Washington, George, 1732-1799
Language: English
Description: Dedicated to Martha Washington
Link: http://opac.newsbank.com/select/evans/38998

The resignation of His Excellency George Washington, president of the United States; and his
Genre: Document
Creator: Washington, George, 1732-1799
Date: 1796
Published in: United States--New York--New York
Publisher: James Oram, by, 1796
Language: English
Description: Signed on p. 26: G. Washington. United States, 17th September, 1796
Link: http://opac.newsbank.com/select/evans/31535

President Washington's resignation, and address to the citizens of the United States, September 17
Creator: Washington, George, 1732-1799
Date: 1796
Published in: United States--Massachusetts--Newburyport
Geocode: lat=42.8125 lng=-70.877778
Publisher: Barrett, William, 1796
Language: English

(asc | desc) Date
(asc | desc) Creator
(asc | desc) Score

collapse or expand all

Subject facet
- Washington, George, 1732-1799 (37)
- Death and burial (24)
- Washington, George, 1732-1799, Death and burial (24)
- More...

Person facet
- Washington, George, 1732-1799 (7)
- Paine, Thomas, 1737-1809 (2)
- Searson, John (2)
- More...

Organization facet
- United States. President (1789-1797 : Washington) (5)
- United States. Army (2)
- United States. Continental Congress (2)
- More...

Publisher facet
- s.n, 1800 (4)
- Green, Samuel, 1800 (2)

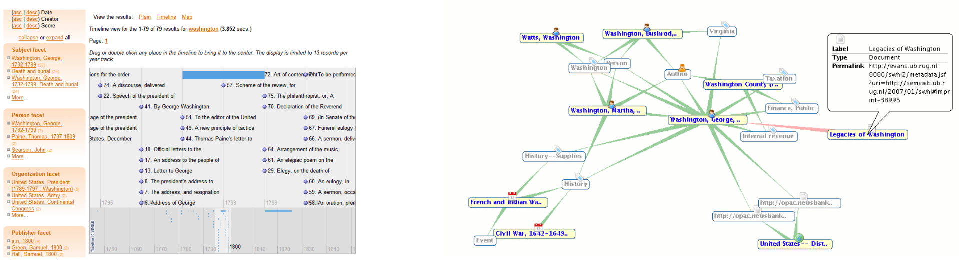
View the results: Plain Timeline Map

Map view for the **1-79** of **79** results for washington (**0.054** secs.)

Page: 1

The map is limited to 100 records per view. Only records with location data will be displayed.

Figure 5.2: Search result pages for 'George Washington' on the current SWHi prototype. The search results can be displayed in multiple views, including a textual view (left), and a Map view (right).


(asc | desc) Date
(asc | desc) Creator
(asc | desc) Score

collapse or expand all

Subject facet
- Washington, George, 1732-1799 (37)
- Death and burial (24)
- Washington, George, 1732-1799, Death and burial (24)
- More...

Person facet
- Washington, George, 1732-1799 (7)
- Paine, Thomas, 1737-1809 (2)
- Searson, John (2)
- More...

Organization facet
- United States. President (1789-1797 : Washington) (5)
- United States. Army (2)
- United States. Continental Congress (2)
- More...

Publisher facet
- s.n, 1800 (4)
- Green, Samuel, 1800 (2)
- Hall, Samuel, 1800 (2)

View the results: Plain Timeline Map

Timeline view for the **1-79** of **79** results for washington (**3.852** secs.)

Page: 1

Drag or double click any place in the timeline to bring it to the center. The display is limited to 13 records per year track.

ions for the order
72. Art of contentment
71. To be performed
74. A discourse, delivered
57. Scheme of the review, for
22. Speech of the president of
75. The philanthropist: or, A
41. By George Washington,
70. Declaration of the Reverend
age of the president
54. To the editor of the United
69. (In Senate of the
age of the president
49. A new principle of tactics
67. Funeral eulogy
States. December
44. Thomas Paine's letter to
66. A sermon, deliv
18. Official letters to the
64. Arrangement of the music,
17. An address to the people of
61. An elegiac poem on the
13. Letter to George
29. Elegy, on the death of
8. The president's address to
60. An eulogy, in
7. The address, and resignation
59. A sermon, occa
6. Address of George
58. An oration, pron
1795
1798
1799
Timeline © SIMILE
1750
1760
1770
1780
1790
1800
1810
1820
1830
1840

Label	Legacies of Washington
Type	Document
Permalink	http://evans.ub.rug.nl:8080/swhi2/metadata.jsf?uri=http://semweb.ub.rug.nl/2007/01/swhi#Imprint-38995

Figure 5.3: Search result pages for 'George Washington' on the current SWHi prototype. The search results can be displayed in multiple views, including a Timeline view (left), and a contextual Network view (right).

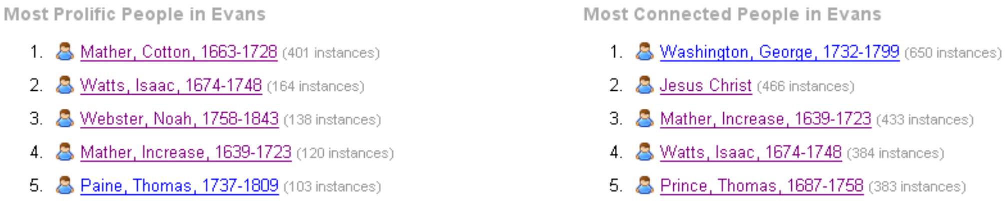
2. *The user identifies an interesting piece of information, or 'berry'*, and adds the berry to the basket (see figure 5.7). Every new berry helps the user contextualize the information within his research domain.

3. *The system updates the berry basket.* The list of semantic relations and properties, shared by the items in the basket, is also updated. These shared relations and properties are used to compute 'semantically similar' items, based on the items the historian has deemed of interest to his research purposes.

4. *The berry basket detail ('checkout') page can be displayed on demand*, either during or at the end of the information-seeking process. (see figure 5.8)

5. (a) The user sees the list of semantically similar—and potentially relevant—items, and decides to add some more items to the basket, and resumes the information-seeking process (step 1), or:

   (b) The user reviews the items he has collected in his basket, decides he has fully developed the contextual overview of the research domain he desired, and starts the next research phase (writing, or performing other research tasks). (End of iteration);



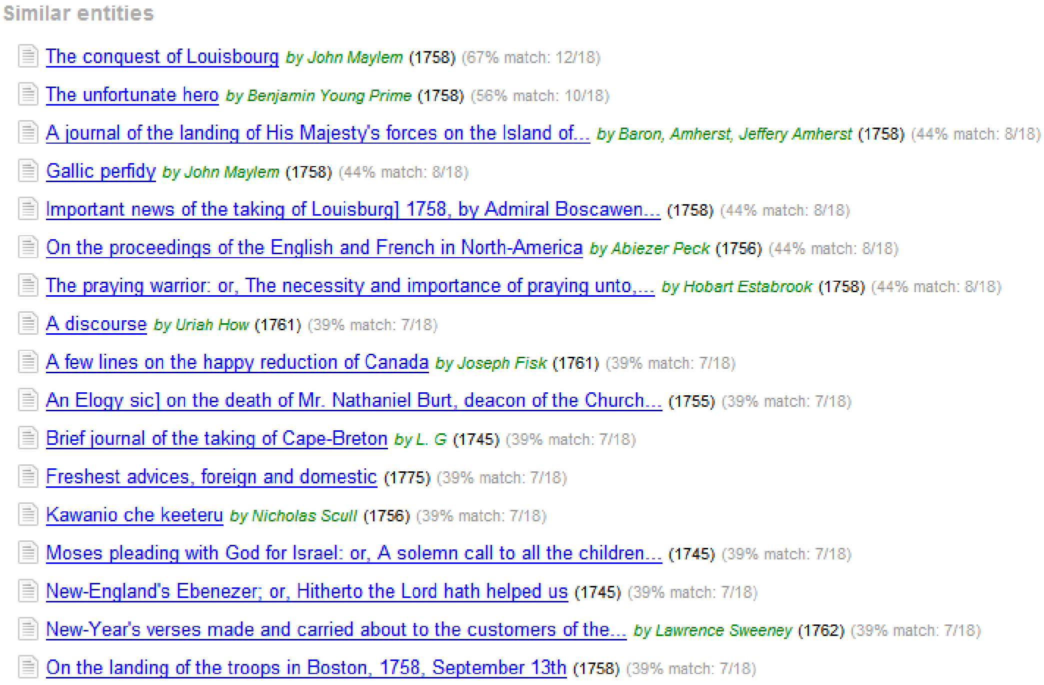
Most Prolific People in Evans

1. Mather, Cotton, 1663-1728 (401 instances)
2. Watts, Isaac, 1674-1748 (164 instances)
3. Webster, Noah, 1758-1843 (138 instances)
4. Mather, Increase, 1639-1723 (120 instances)
5. Paine, Thomas, 1737-1809 (103 instances)

Most Connected People in Evans

1. Washington, George, 1732-1799 (650 instances)
2. Jesus Christ (466 instances)
3. Mather, Increase, 1639-1723 (433 instances)
4. Watts, Isaac, 1674-1748 (384 instances)
5. Prince, Thomas, 1687-1758 (383 instances)

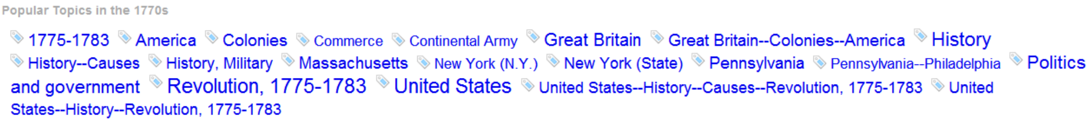
Figure 5.4: Two examples of ranked textual lists, denoting people of interest in the SWHi corpus. These lists are displayed on the 'home page' (main overview page) of the SWHi electronic finding aid website. The list on the left is the list of authors with the most imprints in the collection; the list on the right denotes the persons with the most semantic connections in the ontology, the spiders in the Semantic Web, so to speak.



Similar entities

- The conquest of Louisbourg by John Maylem (1758) (67% match: 12/18)
- The unfortunate hero by Benjamin Young Prime (1758) (56% match: 10/18)
- A journal of the landing of His Majesty's forces on the Island of... by Baron, Amherst, Jeffery Amherst (1758) (44% match: 8/18)
- Gallic perfidy by John Maylem (1758) (44% match: 8/18)
- Important news of the taking of Louisburg] 1758, by Admiral Boscawen... (1758) (44% match: 8/18)
- On the proceedings of the English and French in North-America by Abiezer Peck (1756) (44% match: 8/18)
- The praying warrior: or, The necessity and importance of praying unto,... by Hobart Estabrook (1758) (44% match: 8/18)
- A discourse by Uriah How (1761) (39% match: 7/18)
- A few lines on the happy reduction of Canada by Joseph Fisk (1761) (39% match: 7/18)
- An Elogy sic] on the death of Mr. Nathaniel Burt, deacon of the Church... (1755) (39% match: 7/18)
- Brief journal of the taking of Cape-Breton by L. G (1745) (39% match: 7/18)
- Freshest advices, foreign and domestic (1775) (39% match: 7/18)
- Kawanio che keeteru by Nicholas Scull (1756) (39% match: 7/18)
- Moses pleading with God for Israel: or, A solemn call to all the children... (1745) (39% match: 7/18)
- New-England's Ebenezer; or, Hitherto the Lord hath helped us (1745) (39% match: 7/18)
- New-Year's verses made and carried about to the customers of the... by Lawrence Sweeney (1762) (39% match: 7/18)
- On the landing of the troops in Boston, 1758, September 13th (1758) (39% match: 7/18)

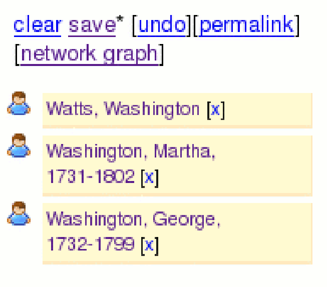
Figure 5.5: Suggested similar items for Imprint #14254, *The Conquest of Louisburg*, based on the commonly shared characteristics discussed in section 5.1.1. The closest document match has a 67% similarity to this imprint (sharing 12 out of a total of 18 characteristics). Not coincidentally, the closest match is an earlier imprint of the same work. Differences in publication year, publisher, and other metadata characteristics account for the fact that the match in this case is not closer to 100%.



Popular Topics in the 1770s

1775-1783 America Colonies Commerce Continental Army Great Britain Great Britain--Colonies--America History History--Causes History, Military Massachusetts New York (N.Y.) New York (State) Pennsylvania Pennsylvania--Philadelphia Politics and government Revolution, 1775-1783 United States United States--History--Causes--Revolution, 1775-1783 United States--History--Revolution, 1775-1783

Figure 5.6: A thematic cloud denoting the 'zeitgeist' of the 1770s. This cloud provides a quick insight into the most important (most frequent) themes of this decade. Here we see a textual visualization implemented on a subcategory level (Chronological, By Decade, 1770).



clear save* [undo][permalink]
[network graph]

Watts, Washington [x]

Washington, Martha, 1731-1802 [x]

Washington, George, 1732-1799 [x]

Figure 5.7: The contents of the berry basket, as it is displayed inline, as part of each page in the electronic finding aid (detail). In this semantic berry basket, historians are not only able to add persons of interest, but also all documents, events, places, etc. they deem interesting.

5.2.1 IV and the Berry Basket

The Berry Basket contains all the items the user has deemed interesting enough to set aside, to save and to examine further at a later stage. Normally, in a shopping basket, one examines the basket's contents, which are presented solely in a textual form.

In this Berry Basket model, an important role is reserved for Information Visualization. The use of fast-response, interactive web pages and Information Visualizations, as Shneiderman (1996) proposes, will allow for fast manipulation. It will enable the user to zoom into and find items of interest in ways that are impossible in a traditional, static finding aid. All features of the system are directed toward attaining the same goals: assisting the user by amplifying cognition and assisting information-seeking and -discovery.

By pointing out items in the collection to the user, and providing different view points to find items of interest, the system already helps the user a great deal. But there is an even further level, an extra dimension which can even amplify amplify the user's cognitive abilities even further: these same methods—view-transformation and similarity-suggestion—can also applied to the contents of the Berry Basket. Figure 5.7 shows the contents of the berry basket displayed in a Clustermap view.

On each page and on each level, the interactive Information Visualizations (textual, geographical, temporal, and contextual) provide the user with a full array of options and views to see the information in context. The various views of this semantically enhanced berry basket all have the following options:

- Manipulate the entity/berry basket/list (add/delete/promote) and/or its set of (shared) relations/properties.

- View the entity list in context, to relate them to other entities and to discern patterns.

- View a list of suggested entities (persons of interest, documents, topics, etc.) based on the predictive properties of shared semantic relations, and the ability to add them to the existing set (baskets), while refining and redefining the schema/context mapping even further.

The same view transformations that are available in the searching and browsing interface—at the information-gathering stage—can also be applied within the berry basket itself. The user can examine the information gathered in the basket so far, and contextualize it 'internally' in his mental schema by displaying it in multiple dimensions (textual, temporal, geographical/spatial, and contextual/networked).

Just as a set of suggested similar items can be computed for a single entity in the ontology, it can also be computed for a set consisting of multiple items. Accordingly, we can also generate a list of items that share common characteristics with the contents of the berry basket as a whole, from the collective aggregated properties of *the entire set*. The items suggested in this process are the ones that fit best with the berries that were already picked and placed in the basket.[11]

By providing these advanced contextualization options the system assists the user in developing his mental schema on yet a further level.[12]

5.3 Chapter Summary

The automatic suggestion of similar or related items is not new in web-based information systems. Retailers have been tracking their customers' behavior (on- and off-line) for ages, with the intent to target these customers with 'personal' offers they might be interested in, considering their past buying habits. Offering the user a personalized on-line experience, in which items that might be of interest to them are delivered to the user's doorstep, has proven to be beneficial to both parties. For some reason this user-experience-enhancing practice has not gained much ground in scholarly and bibliographic environments, although an on-line bibliographic finding aid is not dissimilar to an on-line bookstore. The bibliographic metadata we have at our disposal is rich enough to be able to provide our users with a list of items that is potentially of interest to them. The simple three-part semantic data structure of a triple proves to be a very powerful one in this respect. If we combine our user-centered outlook with this powerful data structure, and also add the cognition-augmenting benefits of Information Visualization to the picture, we get a powerful mixture that can assist historians in their information-seeking practices in novel ways, and empower them to

[11] One might say that the system at this moment has discovered what the historians' favorite kind of berries are.

[12] On a side note, the different IV views of the Berry Basket also has another, although secondary potential use, in that it doubles as an interactive on-line map-building, timeline-generating, and graph-making environment: by adding the items needed into the basket, and subsequently merely switching views, one can easily build a desired visualization to include in one's research paper or on-line publication.

form ‘the information in their heads’ in many respects. We like to think that the Semantically-enhanced Berry Basket is more than merely the sum of its parts.

Chapter 6

Conclusion

From the start of the research for this thesis, we have attempted to employ a multi-faceted approach to the problem at hand. Whilst employing a user-centered approach, we have tried to incorporate and combine methods and insights from many different fields, from the source-oriented bibliographical research of print history and library science, to the realms of Human-Computer Interaction, Usability Engineering, Information Visualization, Information Retrieval and the Semantic Web.

The outcome of this broad approach—that all these seemingly divergent roads would ultimately lead us in the same direction, augmenting cognition—has even managed to surprise ourselves. An ideal on-line electronic finding aid for bibliographical data should not merely make the data available and searchable, but should set out to assist the user in every way possible. The user in question is the historian, and we have found his information-seeking behaviour to differ slightly from what we had expected, to differ from the omnipresent traditional one–stop Information Retrieval model of 'one–query/one–use'. Perhaps, if we had known this from the outset, we would have chosen a different audience of scholarly users, preferable one that would display a behavior that conformed slightly closer to this traditional model.

Instead, we have found that historians display a seemingly more irrational behavior, centered around serendipitous item and pattern discovery. Although they display a traditional initial distrust towards new technologies, historians have proven to be very adaptive in accomodating their information-seeking behavior to the nature of the their resource. This adaptive trait is perhaps best exemplified by historians' habits of collecting names and adhering to the provenance method, which they derived from the traditional organization of most archives, which were set up around provenance and authorship.

Historians seem to display behavior that can best be explained by Marcia Bates' Berrypicking Model (Bates, 1989). In this model, the information-seeker sets out with an open-ended search goal, in which a newly discovered interesting piece of information (the 'berry') is 'picked', put into a 'basket', which signifies an mental contextual model of the domain of interest. Every new berry the information-seeker places into the basket adds a new piece to the contextual puzzle, which causes the information-seeker to continuously re-evaluate and shift the information-seeking trail through the information forest: every new piece of information the information-seeker uncovers further adds to the 'information in his head', in turn further amplifying the information-seekers internal cognitive process. An ideal electronic finding aid for historical primary resources should accomodate such a behaviour and stimulate the information-seekers own quest for building his contextual domain knowledge. The berrypicking model bears some striking similarities to a the organization of an on-line retail store, in which a user shops around for interesting buys, adds them to the shopping basket, and ultimately checks the items out when his shopping urges are satisfied. This analogy with an on-line shop is further exemplified by the 'Name-Drawer Schema' Charles Cole envisioned in order to translate the historian's information-behavior to an electronic bibliographic interface.

Furthermore, we have found that translating such an interface to an on-line environment offers interesting new ways to allow for pattern discovery and serendipitous information-seeking. Adding Information Visualizations like interactive and descriptive maps and timelines to the electronic finding aid's interface further improves its potential to augment cognition.

Along with a user-centered approach we have also employed a data-centered approach to investigate the intricacies of the *Evans* collection of Early American Imprints, of the data source chosen by the SWHi project to serve as the basis for the Bibliographic Semantic Web project. This approach has provided a unique insight into the use of bibliographic (meta-)data by several generations of historians, bibliographers, and librarians. We have identified several pivotal points and objects around which historical information is structured: besides documents (which is the traditional focus of bibliographic metadata), historical information centers around agents (persons and organizations), and can be further extended into temporal and geographical dimensions. This multi-object semantic data taxonomy fits well with recent initiatives to move away from document-centered metadata, like FRBR. By understanding both resource and user one can get a step closer to the creation and organization of an effective, efficient and usable resource for new generations of users.

These combined insights have led to a variation on Cole's Name-Drawer Schema and Bates' Berrypicking model, which we have dubbed the 'Semantically-enhanced Berry Basket'. The proposed model and the algorithm for automatic suggestion of 'new berries' will—when fully applied in a next stage of the SWHi project—facilitate and enhance the historian's information-seeking and internal 'context-building' goals, by augmenting cognition and pattern-recognition. This model expands on Cole's and Bates' ideas, and applies them in a modern-day information environment. It uses several techniques which, although they are becoming seemingly common on todays web, were not yet applied to scholarly environments, certainly not to the extent as has been shown here. The Semantically-enhanced berry basket builds on three powerful pillars which all—in the previous sections—have been shown to display traits that are beneficial to a user-centered on-line electronic finding aid: historians' information-seeking behavior, Information Visualization, and the Semantic Web.

6.1 Discussion: A Critical Note

Although historians are avid users of libraries, archives, and bibliographical collections, this does not necessarily make them the ideal users of *electronic* versions of these information resources. Historians are researchers of the past. With their attention directed towards the past, it is perhaps not strange that historians are usually not among the first to look at the future as the early adopters of new technologies and new research tools.

While historical evidence can be found in—mostly written—primary sources, historians are primarily text-oriented, and moreover they a bit reluctant to use—or perhaps just reluctant to admit the usefulness of—computers as their most important scholarly tools of the trade. This preference for textual sources and 'analog' tools has been formed by a research tradition which spans centuries. A change of customs will therefore be relatively slow: it will take some time before digital tools or digital versions of resources will earn the same level of recognition and gain the same level of trust as written or analog resources: the documents historians research have been around for centuries: in the mind of a historians, computer-assisted research and digital tools have been around for just an instant.

If we look at Information Visualization and how it is used, we have on one hand our users, historians, who 'typically scoff at a visual display as mere decoration' (Staley, 2003). On the other hand we have system developers, a technology-oriented species who sometimes let technology lead the way. Often, with a website or an interactive environment, it is all too tempting to let this happen—to let technology lead—and with this, along the way, the user-centeredness gives way. It happens a lot. It has happened to myself more often than I care to admit.

Effective visualizations and interfaces should have *users—and usage*—as their main interest, and this can only be achieved by keeping clarity and simplicity in mind when one develops and deploys (interactive) information visualizations. Using high-end, high-tech abstract visualizations like Clustermaps and three-dimensional conceptual models of information spaces—how appealing they may be to technologists and developers—may not fare well with users that set out to find information, who prefer text, along with simple yet illustrative information visualizations, as most humanists seem to.

If even seasoned Information and Library Science professionals have trouble grasping and understanding these kinds of visualizations, one should thoroughly question whether these visualizations should be used as a way to let people navigate this information space. For those few wishing to explore and experiment with advanced: please do so, but as these IVs haven't well attained mainstream acceptance and usage yet, it surely does not need to be positioned as anything more than a peripheral, experimental IR tool.

Visual/clustered Information Retrieval interfaces should, at this time, still be considered a mere novelty (although an interesting and promising one). For a research field like history, where adaptation of new technologies is not first thing at the scholars' minds, it might be a wise choice to stick with information aids that stay closer to current research methodology, closer to their information–seeking behaviour. If users cannot instantly familiarize and identify themselves with the first tastes they have of new technologies, they will discard and subsequently disregards these 'novel' technologies.

Still, I think it is good to have these 'advanced' Information Visualizations available for experimenting and for use among computer scientists and other proponents of these technologies. Maybe, one day, some of the visualization methods we have been experimenting with *will* enter the main stream. But with not-so-technologically-savvy users like humanists, one should not go overboard with trying to incorporate every bit of new technology just for the sake of it. I tend to agree with David Staley when it comes to the role of visual displays of information. However, visual displays *can* be highly effective even as a mere illustration. They do not necessarily need to be put in the center of attention: a place in the shadows, at the periphery will do for now.

6.2 Recommendations and Future Work

The development of an on-line information-seeking tool is a neverending process: technology progresses rapidly, insights change, and user's expectations of on-line services are also subject to change. Evaluating the product with users forms an integral part of the user-centered design process. Most of the concepts and ideas brought forward in this thesis have not been sufficiently tested with actual users—i.e. historians—yet.

As far as our ontology is concerned, it can be further expanded, and also improved in some places (events and printers come to mind here as entity types to focus on. Furthermore, we can extract even more meaningful relations from our metadata, especially where the interpersonal (knows) relations are concerned. In the near future, the SWHi project will have access to the actual full-text corpus and to the actual scanned versions of the *Evans* collection, which will provide us with additional opportunities to extract semantic relations. It also makes further improvements with respect to full-text Information Retrieval and interactive page-browsing interfaces well within reach of this project.

Other interesting research and experimentation opportonities lie in adding community-powered Social (Web 2.0/Library 2.0) features like bookmark sharing, community-based recommendation, tagging, to our already user-centered Semantic application. Scholarly researchers should also benefit from other opportunities Web 2.0 and Library 2.0 applications offer to interlink and make use of each others resources. The SWHi project's electronic finding aid should be linked to other bibliographic resources, like the University of Groningen's own Livetrix system, to maximize the use of the primary sources by connecting them directly to the scholarly articles and secondary sources that are based directly on these primary resources.

Also, further experimentation an refinement of the interface elements should take place, especially where semantic similarity and advanced weighting mechanisms are concerned. The same goes for further research into effective interactive visualisations of (semantic) historical data.

Lastly, we could further improve, digitize and transcribe the full-text data corpus of *Evans* by joining the Text Creation Partnership's initiative. The eighty-odd Dutch language imprints in Evans can be an interesting starting point to ensure that *Evans* collection will be used by generations of historians to come.

Appendix A

Works Cited

Bibliography

Ankolekar, A., Krötzsch, M., Tran, T., and Vrandecic, D. (2007). The two cultures: mashing up web 2.0 and the semantic web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 825–834, New York, NY, USA. ACM Press.

Arms, W. Y. (2000). *Digital Libraries.* MIT Press, Cambridge, MA, USA.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modeling. In Baeza-Yates, R. and Ribeiro-Neto, B., editors, *Modern Information Retrieval*, pages 19–72. ACM Press, New York, NY, USA.

Bass, R. and Rosenzweig, R. (2001). Rewiring the history and social studies classroom: Needs, frameworks, dangers, and proposals. *Boston University Journal of Education*, 181(3):41–61.

Bates, M. J. (1989). The design for browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–431.

Bates, M. J. (2002). Speculations on browsing, directed searching, and linking in relation to the bradford distribution. In Bruce, H., Fidel, R., Ingwersen, P., and Vakkari, P., editors, *Emerging Frameworks and Methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)*, pages 137–150, Greenwood Village, CO, USA. Libraries Unlimited.

Bates, M. J. (2005a). Berrypicking. In Fisher, K. E., Erdelez, S., and McKechnie, L., editors, *Theories of information behavior*, pages 58–62. American Society for Information Science and Technology, Medford, NJ, USA.

Bates, M. J. (2005b). An introduction to metatheories, theories, and models. In Fisher, K. E., Erdelez, S., and McKechnie, L., editors, *Theories of information behavior*, pages 1–24. American Society for Information Science and Technology, Medford, NJ, USA.

Beattie, D. L. (1990). An archival user study: Researchers in the field of women's history. *Archivaria*, 29:33–50.

Berners-Lee, T. (2003). Foreword. In Fensel, D., Hendler, J. A., Lieberman, H., and Wahlster, W., editors, *Spinning the Semantic Web. Bringing the World Wide Web to Its Full Potential*, pages xi–xxiii. MIT Press, Cambridge, MA, USA.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *The Scientific American*. `http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21`, visited March 21, 2007.

Bristol, R. (1970). *Supplement to Charles Evans' American bibliography*. Bibliographical Society of America and the Bibliographical Society of the University of Virginia, Charlottesville, VA, USA.

Card, S. (2003). Information visualization. In Jacko, J. A. and Sears, A., editors, *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, pages 544–582. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Card, S., Mackinlay, J., and Shneiderman, B. (1999). *Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, CA, USA.

Case, D. O. (1991a). The collection and use of information by some american historians: A study of motives and methods. *Library Quarterly*, 61(1):61–82.

Case, D. O. (1991b). Conceptual organization and retrieval of text by historians: The role of memory and metaphor. *Journal of the American Society for Information Science*, 42(9):657–668.

Cole, C. (2000a). Inducing expertise in history doctoral students via information retrieval design. *Library Quarterly*, 70(1):86–109.

Cole, C. (2000b). Name collection by ph.d. history students: Inducing expertise. *Journal of the American Society for Information Science*, 51(5):444–455.

Delgadillo, R. and Lynch, B. P. (1999). Future historians: Their quest for information. *College & Research Libraries*, 60(3):245–259.

Duff, W. M. and Johnson, C. A. (2002). Accidentally found on purpose: Information-seeking behaviour of historians in archives. *Library Quarterly*, 72(4):472–496.

Evans, C. T. (1903–1935). *American bibliography: a chronological dictionary of all books, pamphlets, and periodical publications printed in the United States of America from the genesis of printing in 1639 down to and including the year 1820 : with bibliographical and biographical notes.* Various Publishers. 13 volumes.

Fahmi, I., Zhang, J., Ellerman, H., and Bouma, G. (2007). Swhi system description: A case study in information retrieval, inference, and visualization in the semantic web. In *ESWC 2007: Proceedings of the European Semantic Web Conference*, Innsbruck, Austria. Springer. `http://www.eswc2007.org/pdf/eswc07-fahmi.pdf`.

Faulkner, X. (2000). *Usability Engineering.* Palgrave, Basingstoke, UK.

Finin, T., Mayfield, J., Joshi, A., Cost, R. S., and Fink, C. (2005). Information retrieval and the semantic web. In *HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005*, pages 113a–113a.

Fisher, K. E., Erdelez, S., and McKechnie, L. (2005). Preface. In *Theories of information behavior*, pages xix–xxii. American Society for Information Science and Technology, Medford, NJ, USA.

Fluit, C., Sabou, M., and van Harmelen, F. (2002). Ontology-based information visualization. In *Visualizing the Semantic Web*, pages 36–48. Springer, London, UK.

Fox, E. A. and Sornil, O. (1999). Digital libraries. In Baeza-Yates, R. and Ribeiro-Neto, B., editors, *Modern Information Retrieval*, pages 415–432. ACM Press, New York, NY, USA.

Geroimenko, V. and Chen, C. (2002). *Visualizing the Semantic Web. XML-based Internet and Information Visualization.* Springer, London, UK.

Hacknos, J. T. and Redish, J. C. (1998). *User and Task Analysis for Interface Design.* Wiley, New York, NY, USA.

Halvey, M. J. and Keane, M. T. (2007). An assessment of tag presentation techniques. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1313–1314, New York, NY, USA. ACM Press.

Hasan-Montero, Y. and Herrero-Solana, V. (2006). Improving tag-clouds as a visual information retrieval interfaces. In *InSciT2006. Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies.*

Hearst, M. A. (1999). User interfaces and visualization. In Baeza-Yates, R. and Ribeiro-Neto, B., editors, *Modern Information Retrieval*, pages 257–324. ACM Press, New York, NY, USA.

Herzog, C., Luger, M., and Herzog, M. (2007). Combining social and semantic metadata for search in a document repository. In *Bridging the Gep between Semantic Web and Web 2.0 (SemNet 2007)*, pages 14–21.

Holley, E. G. (1963). *Charles Evans: American bibliographer.* University of Illinois Press.

Huynh, D. F., Karger, D. R., and Miller, R. C. (2007). Exhibit: lightweight structured data publishing. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 737–746, New York, NY, USA. ACM Press.

IFLA (1998). Functional requirements of bibliographic records: final report. Technical report, IFLA Study Group on the Functional Requirements of Bibliographic Records. `http://www.ifla.org/VII/s13/frbr/frbr.pdf`, visited April 10, 2007.

Khare, R. and Çelik, T. (2006). Microformats: a pragmatic path to the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 865–866, New York, NY, USA. ACM Press.

Kimani, S., Catarci, T., and Cruz, I. F. (2002). Web rendering systems: Techniques, classification criteria and challenges. In *Visualizing the Semantic Web*, pages 63–89. Springer, London, UK.

Krummel, D. W. (2005). Early american imprint bibliography and its stories: An introductory course in bibliographical civics. *Libraries & Culture*, 40(3):239–250.

Li, R., Bao, S., Yu, Y., Fei, B., and Su, Z. (2007). Towards effective browsing of large scale social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 943–952, New York, NY, USA. ACM Press.

MARC (2006). Functional analysis of marc 21. Technical report, MARC Standards Office, Library of Congress. `http://www.loc.gov/marc/marc-functional-analysis/frbr.html`, visited July 20, 2007.

Nielsen, J. (1993). *Usability Engineering.* AP Professional, Boston, MA, USA.

Nielsen, J. (2000). *Designing Web Usability: The Practice of Simplicity.* New Riders, Indianapolis, IN, USA.

Nielsen, J. (2003). Usability 101: Fundamentals and design. *Useit.com.* `http://www.useit.com/alertbox/20040825.html`, visited June 1, 2007.

Norman, D. A. (1988). *The Design of Everyday Things.* Basic Books, New York, NY, USA. Originally published as 'The Psychology of Everyday Things'.

Orbach, B. C. (1991). The view from the researcher's desk: Historians' perceptions of research and repositories. *American Archivist*, 54(1):28–43.

Pandit, S. and Olston, C. (2007). Navigation aided retrieval. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 391–400, New York, NY, USA. ACM Press.

Pedersen, G. S. (1993). A browser for bibliographic information retrieval, based on an application of lattice theory. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 270–279, New York, NY, USA. ACM Press.

Rasmussen, E. M. (1999). Libraries and bibliographical systems. In Baeza-Yates, R. and Ribeiro-Neto, B., editors, *Modern Information Retrieval*, pages 397–413. ACM Press, New York, NY, USA.

Reese, W. S. (1990). *The First Hundred Years of Printing in British North America: Printers and Collectors.* American Antiquarian Society, Worcester, MA, USA. `http://www.reeseco.com/papers/first100.htm`, visited June 27, 2007.

Riley, S. T. (1971). Review of guide to the study of united states imprints. by g. thomas tanselle. *New England Quarterly*, 45(3):460–462.

Rivadeneira, A. W., Gruen, D. M., Muller, M. J., and Millen, D. R. (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 995–998, New York, NY, USA. ACM Press.

Rouse, W. B. and Rouse, S. H. (1984). Human information seeking and design of information systems. *Information Processing & Management*, 20(1–2):129–138.

Shipton, C. K. (1955). *The American bibliography of Charles Evans: a chronological dictionary of all books, pamphlets, and periodical publications printed in the United States of America from the genesis of printing in 1639 down to and including the year 1800 : with bibliographical and biographical notes. Volume 13. 1799-1800.* American Antiquarian Society.

Shipton, C. K. and Mooney, J. E. (1969). *National Index of American Imprints through 1800: The short-title Evans.* American Antiquarian Society, Worchester, MA, USA. 2 volumes.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, page 336, Washington, DC, USA. IEEE Computer Society.

Shneiderman, B. (1997). Designing information-abundant web sites: issues and recommendations. *International Journal of Human-Computer Studies*, 47(1):5–29.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Addison-Wesley, Reading, MA, USA.

Staley, D. J. (2003). *Computers, Visualization and History. How New Technology Will Transform Our Understanding of the Past.* American Association for History and Computing, Armonk, NY, USA.

Stieg-Dalton, M. and Charnigo, L. (2004). Historians and their information sources. *College & Research Libraries*, 65(5):400–425.

Tanselle, G. T. (1971). *Guide to the study of United States imprints.* Harvard University Press, Cambridge, MA, USA.

Tibbo, H. R. (2002). Primarily history: historians and the search for primary source materials. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 1–10, New York, NY, USA. ACM Press.

Tufte, E. R. (1983). *The visual display of quantitative information.* Graphics Press, Cheshire, CT, USA.

Tvarozek, M. and Bielikova, M. (2007). Adaptive faceted browser for navigation in open information spaces. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1311–1312, New York, NY, USA. ACM Press.

W3C (2001). W3c semantic web frequently asked questions. Technical report, World Wide Web Consortium. `http://www.w3.org/2001/sw/SW-FAQ#What1`, visited June 28, 2007.

Welling, G. M. (1998). *The Prize of Neutrality. Trade Relations between Amsterdam and North America 1771-1817. A Study in Computational History.* Amsterdamse Historische Reeks, Amsterdam, NL.

Yakel, E. (2005). Archival intelligence. In Fisher, K. E., Erdelez, S., and McKechnie, L., editors, *Theories of information behavior*, pages 49–53. American Society for Information Science and Technology, Medford, NJ, USA.

Yakel, E. and Torres, D. A. (2003). Ai: Archival intelligence and user expertise. *American Archivist*, 66(1):51–78.

Zhang, J. (2006). Mapping metadata is more. making more use out of existing metadata from digital libraries using semantic web technologies. Master's thesis, Rijksuniversiteit Groningen, the Netherlands. `http://www.let.rug.nl/alfa/scripties/JunteZhang.html`.